https://doi.org/10.15388/vu.thesis.923
https://orcid.org/0009-0004-8815-2478

VILNIUS UNIVERSITY

Modestas Motiejauskas

Convolutional Neural Network Architectures and Training Improvements: Visual Emotion Recognition Case

DOCTORAL DISSERTATION

Natural Sciences,
Informatics (N 009)
VILNIUS 2026

This dissertation was prepared between 2021 and 2025 at Vilnius University.

Academic supervisor – Prof. Habil. Dr. Gintautas Dzemyda (Vilnius University, Natural Sciences, Informatics – N 009).

This Doctoral Dissertation will be Defended in a Public Meeting of the Dissertation Defence Panel:
Chair – Prof. Dr. Olga Kurasova (Vilnius University, Natural Sciences, Informatics – N 009).
Members:
Prof. Dr. Romas Baronas (Vilnius University, Natural Sciences, Informatics – N 009),
Prof. Dr. Diana Kalibatienė (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – T 007),
Dr. Gerda Ana Melnik-Leroy (Vilnius University, Natural Sciences, Informatics – N 009),
Prof. Dr. Audris Mockus (University of Tennessee, USA, Natural Sciences, Informatics – N 009).

The dissertation will be defended at a public meeting of the Dissertation Defence Panel at 2 p.m. on 25th of May in the Conference Room D1-16.1 of the Medical Science Centre of Vilnius University.
Address: Žaliųjų ežerų St. 2, LT-08406, Vilnius, Lithuania.

https://doi.org/10.15388/vu.thesis.923
https://orcid.org/0009-0004-8815-2478

VILNIAUS UNIVERSITETAS

Modestas Motiejauskas

Konvoliucinių neuroninių tinklų architektūrų ir mokymo gerinimas: emocijų atpažinimo vaizduose atvejis

DAKTARO DISERTACIJA

Gamtos mokslai,
Informatika (N 009)
Vilnius

VILNIUS 2026

Disertacija rengta 2021–2025 metais Vilniaus universitete.

Mokslinis vadovas – prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, gamtos mokslai, informatika – N 009).

Gynimo taryba:
Pirmininkė – prof. dr. Olga Kurasova (Vilniaus universitetas, gamtos mokslai, informatika – N 009).
Nariai:
prof. dr. Romas Baronas (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
prof. dr. Diana Kalibatienė (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – T 007),
dr. Gerda Ana Melnik-Leroy (Vilniaus universitetas, gamtos mokslai, informatika – N 009),
prof. dr. Audris Mockus (Tenesio universitetas, JAV, gamtos mokslai, informatika – N 009).

Disertacija ginama viešame Gynimo tarybos posėdyje 2026 m. gegužės 25 d. (pirmadienis) 14 val. Vilniaus universiteto Medicinos mokslo centro D1-16.1 konferencijų salėje.
Adresas: Žaliųjų ežerų g. 2, LT-08406, Vilnius, Lietuva.

ACKNOWLEDGMENTS

I express my gratitude to my supervisor Prof. Habil. Dr. Gintautas Dzemyda for their patience, help and guidance during the whole preparation of this dissertation. I would also like to thank the reviewers Prof. Dr. Olga Kurasova and Prof. Dr. Romas Baronas for their valuable time, attention, and constructive comments. I am grateful to my family and loved ones for their patience, and support during all stages of my studies.

ABSTRACT

Visual emotion recognition (VER) is a task in affective computing that aims to automatically recognize emotion in images. In general-purpose images, emotion can be influenced not only on depicted objects but also on color, texture, composition, and scene context. The task is difficult because the same visual elements may be associated with different emotions depending on how they appear in the whole image and how they are perceived by different viewers. In this dissertation, VER is studied in general-purpose images using an eight-category labeling scheme aligned with the employed datasets.

This dissertation proposes a VER model based on a convolutional neural network (CNN). To complement the main visual representation learned by the CNN, Gram matrix modules are added to capture stylistic information from intermediate feature maps, such as texture and color-related patterns. In addition, contrastive-center loss is incorporated into training to improve the separation of emotion classes in the learned feature space and to enhance classification performance on general-purpose visual emotion datasets.

Model performance is evaluated using standard classification metrics, together with feature-space analysis based on dimensionality reduction and clustering quality evaluations. The dissertation also proposes a top-2 cross-sentiment measure for assessing prediction consistency without requiring ground-truth labels. The proposed model, training strategy, and consistency measure contribute to more reliable VER in general-purpose images.

ACRONYMS AND ABBREVIATIONS

AICA	Affective Image Content Analysis
ARI	Adjusted Rand Index
ASR	Ambiguous Sample Ratio
AUC	Area Under the Curve
CES	Categorical Emotion States
CNN	Convolutional Neural Network
DES	Dimensional Emotion Space
FPR	False Positive Rate
NMI	Normalized Mutual Information
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
VEA	Visual Emotion Analysis
VER	Visual Emotion Recognition

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Emotions play a major role in human communication, influencing cognition, decision making, and social interactions. The rapid growth of computational technologies and the dominance of visual media require an understanding of emotional content in images. This has become an important challenge in affective computing – a research field that aims to enable machines to recognize, interpret, and perceive human emotions. In this context, visual emotion recognition (VER) refers to the automatic recognition of emotion in images.

In affective computing, the taxonomy of Mikels et al. [49] comprising eight discrete emotion categories is commonly used to label visual emotion datasets. The taxonomy of Mikels et al defines a practical set of categories widely applied in image emotion research. For this reason, this dissertation primarily follows Mikels' taxonomy, as it matches the emotion labels used in the datasets employed in our experiments.

Research on VER covers several directions. The first focuses on facial expression analysis, where emotions are identified from facial features [33], [32], [86]. A second direction focuses on emotion recognition in general-purpose images, where the affective response depends on global scene composition, color, texture, semantics, or contextual information, rather than on a single facial expression [81], [45]. This dissertation focuses on emotion recognition in general-purpose images because of its broader potential applications.

VER has gained increasing attention because many real-world applications depend on accurately recognizing emotions evoked by images. Automated emotion analysis might support personalized human–computer interaction in mental health (e.g., detecting depressive tendencies from social media images [62]), enable adaptive learning in education (e.g., adjusting lesson pace based on student engagement [26]), and improve social assistive robotics for elderly care [1]. As the spread of the Internet of Things produces an ever-growing volume of visual data, VER has become an important research challenge with wide

applicability.

1.1 RESEARCH PROBLEM

The affective gap, subjectivity in perception, and difficulties in annotating visual emotion images are the main challenges in the field of VER [102]. The affective gap refers to the potential misalignment between the physical features of an image and the emotional state of the viewer. For example, a red rose in a bright background evokes a positive expression, whereas the same physical object in a dark background might evoke a negative emotion. Therefore, the affective gap can be understood as the main problem for VER, since the emotion evoked by an image may depend not only on individual features but also on the image as a whole. Another related difficulty comes from the context aspect. The same image can convey different emotions: a smile in customer service photos might indicate a happy expression, but the environment suggests neutrality or even stress. Viewers may have different emotional reactions to the same image. This can be caused by many personal and contextual factors, such as the cultural background, personality, and social context [99].

Recent advancements in the field of deep learning have led to success in addressing the visual emotion recognition problem. However, the classification results need to be improved, particularly for VER in general-purpose images.

1.2 ACTUALITY

Our surrounding environment is perceived and mediated through visual content. Recent reports estimate that adult Internet users spend an average of 6 hours online every day, with 68% of the global population now connected to the Internet [15]. The vast majority of this content is visual, in the form of images and videos. Visual data represents a medium for expressing and perceiving emotions. This shows a need for automated methods that can effectively interpret large amounts of visual emotion data.

This dissertation addresses these limitations by proposing a CNN-based model and training strategy designed to capture both semantic and stylistic information and improve class separability in the embedding space. The model integrates Gram matrix-based feature representations and is trained using joint optimization with contrastive-center loss. Additionally, the proposed top-2 cross-sentiment measure allows evaluation of the learned feature-space representations without requiring ground-truth labels.

1.3 OBJECT OF THE DISSERTATION

The object of the dissertation is visual images conveying emotions and convolutional neural networks for their analysis.

1.4 GOAL OF THE DISSERTATION

The goal of this dissertation is to develop and evaluate a CNN-based model for visual emotion recognition that captures visual characteristics associated with emotions and improves emotion class separability.

1.5 OBJECTIVES OF THE DISSERTATION

To achieve this goal, the following objectives have been set:

- Conduct a literature review to identify challenges in visual emotion recognition using deep learning methods.
- Conduct a comparative empirical evaluation of selected convolutional neural network architectures to determine the most suitable backbone for visual emotion recognition.
- Develop a CNN-based model for visual emotion recognition in general-purpose images based on the selected architecture, with a focus on improving feature representations to reduce the affective gap.

- Design and validate a CNN-based model integrating contrastive-center loss to improve emotion class separability.

- Analyze the learned feature space using dimensionality reduction, clustering metrics, and the top-2 cross-sentiment measure to evaluate improvements in emotion class separability and representation structure.

1.6 RESEARCH METHODS

This dissertation includes a literature review, the development of a CNN-based model, and experimental and analytical methods for effective VER. The research methods used in this dissertation are described in the following order:

- Theoretical methods:
  Literature review, analysis, and summarizing were used to define the research problem, identify the affective gap as challenge in VER, and establish the novelty of the proposed approach.

- Experimental methods:
  A CNNs-based model was developed.. This included the integration of Gram matrix feature modules, the implementation of contrastive-center loss, and training/evaluation procedures. Dataset pre-processing and augmentation techniques were also applied.

- Analytical methods:
  Model performance was evaluated using classification metrics (accuracy, precision, recall, F1-score) and our proposed top-2 cross-sentiment measure. The structure of the learned feature space was analyzed using dimensionality reduction, clustering, and cluster quality metrics (the adjusted Rand index, normalized mutual information, ambiguous sample ratio) to assess class separability and the quality of feature embedding. Comparative analysis was conducted against the baseline models.

1.7 SCIENTIFIC NOVELTY

This dissertation introduces several novelties to the field of visual emotion analysis. The main contribution of this thesis is the design and evaluation of the proposed CNN-based model for more robust visual emotion classification. This contribution is based on developing a model by integrating additional features to address the affective gap. Additionally, a contrastive-center loss component was proposed to better distinguish visual emotion classes during the model's training process. The following novelties are revealed in this dissertation:

- A novel training optimization strategy is proposed in which the contrastive-center loss is jointly optimized alongside sparse categorical cross-entropy to enhance visual emotion class separability.
- The integration of Gram matrix modules into the convolutional neural network architecture is proposed, aiming to enhance the model with stylistic features.
- A methodology is introduced to analyze the structure of the developed model and to evaluate the effectiveness of integrating the contrastive-center loss into the training process.
- The top-2 cross-sentiment measure is proposed for evaluating internal coherence of model predictions, offering a complementary evaluation without requiring ground-truth labels.

1.8 PRACTICAL VALUE OF THE RESEARCH

The practical value of this dissertation is reflected in the proposed CNN-based model for visual emotion recognition. The integration of Gram matrix modules provides an efficient approach for capturing the stylistic features of images, such as color, texture, and repeating patterns, that are important to convey emotion. This leads to a more reliable emotion classification in diverse and applicable image contexts.

The inclusion of contrastive-center loss in the training process enhances the separation of emotion categories in the feature space, leading

to more reliable predictions and reduced confusion between visually similar classes. These improvements can be directly applied to emotion analysis applications, including content recommendation systems, mental health monitoring, human-computer interaction, and affective multimedia retrieval.

In addition, the experimental framework developed in this dissertation, including dataset preprocessing pipelines, embedding visualization, cluster-based evaluation metrics, and the top-2 cross-sentiment consistency measure, provides a methodology for future studies in affective computing tasks.

1.9 DEFENDED STATEMENTS

- Incorporating Gram matrix modules into the EfficientNetV2S model enables the extraction of additional low-level features, such as color, texture, and repeating patterns. When combined with semantic CNN features, they provide a more complete representation that helps to convey emotion in images.

- The proposed contrastive-center loss increases emotion class separability by compacting intra-class embeddings and enlarging inter-class distances, resulting in improved classification performance compared to standard cross-entropy based loss training.

- Analysis of the learned feature space using dimensionality reduction, clustering metrics, and the top-2 cross-sentiment measure shows that the proposed model learns more coherent and better more clearly separable emotion representations than baseline CNN architectures.

- The top-2 cross-sentiment measure complements accuracy-based evaluation by providing a label-free check of prediction consistency. This rate measures how often the two highest predictions of the model fall into opposite sentiment groups. Adding contrastive-center loss reduces these top-2 cross-sentiment measure cases compared to the baseline.

1.10 APPROBATION OF THE RESEARCH RESULTS

The main results of this dissertation were published in peer-reviewed scientific journals and presented at international and national conferences.

Publications in Clarivate Web of Science journals

[A1] Modestas Motiejauskas, Gintautas Dzemyda (2024). *EfficientNet Convolutional Neural Network with Gram Matrices Modules for Predicting Sadness Emotion*. International Journal of Computers Communications & Control, 19(5), art. no. 6697. https://doi.org/10.15837/ijccc.2024.5.6697

[A2] Modestas Motiejauskas, Gintautas Dzemyda (2025). *The Effective Evaluation of Emotions in the Visual Emotion Images Using Convolutional Neural Networks*. IEEE Access, 13, 139174–139187. https://doi.org/10.1109/ACCESS.2025.3596484

Publications in peer-reviewed international conference proceedings

[B1] Modestas Motiejauskas, Gintautas Dzemyda (2024). *Evaluation of Emotions in Artworks Using EfficientNet Convolutional Network Integrating the Gram Matrix Modules*. In: 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), Xiamen, China, 882–887. https://doi.org/10.1109/ICAIRC64177.2024.10900186

Presentations at international conferences

[C1] Modestas Motiejauskas, Gintautas Dzemyda (2023). *Optimization of EfficientNetV2 Models for Predicting Sadness Emotion*. Numerical Computations: Theory and Algorithms (NUMTA-2023), Calabria, Italy, June 14 - 20, 2023.

[C2] Modestas Motiejauskas, Gintautas Dzemyda (2024). *Evaluation of Emotions in Artworks Using EfficientNet Network Integrating the Gram Matrix Modules*. 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), Xiamen, China, December 27-29, 2024.

Presentations at national conferences

[D1] Modestas Motiejauskas, Gintautas Dzemyda (2022). *On recognizing emotion of sadness in images of a general nature using CNN*. Data Analysis Methods for Software Systems, Druskininkai, Lietuva, December 01-03, 2022.

1.11 OUTLINE OF THE THESIS

This dissertation consists of five primary chapters, followed by conclusions and bibliography.

- Chapter 1 introduces the research topic and provides the background and motivation for visual emotion recognition. It outlines the research problem, objectives, scope, and structure of the dissertation.
- Chapter 2 presents a comprehensive literature review that focuses on the field of visual emotion analysis, related taxonomy models, their applications, and methods addressing the visual emotion recognition problem. This chapter identifies key challenges related to visual emotion analysis and presents a potential state-of-the-art framework based on deep learning.
- Chapter 3 presents the analytical review and experimental comparison of foundational CNNs for visual emotion classification. This chapter covers the models and their suitability analysis for further development of the methodology.

- Chapter 4 presents the proposed methodology for visual emotion analysis using deep learning, focusing on the classification problem, and provides a description of a deep neural network with Gram matrix modules integrated. This chapter also covers the proposed joint contrastive-center loss optimization method.
- Chapter 5 presents the experiments and results retrieved based on the methodology described in a previous chapter. Evaluations and assessments using performance metrics are performed on different visual emotion datasets (FI-8, WEBEmo sadness, EmoSet-118K). Two-dimensional visualizations of the multidimensional data, which describe representations of emotions and which contain embedding features of the network, are also analysed.
- In Conclusions, the key findings of the dissertation are summarized, followed by the Bibliography.

2 LITERATURE REVIEW ON EMOTION RECOGNITION

In this chapter, directions, results, and emerging issues in research on emotion recognition in images are reviewed.

2.1 DESCRIPTION OF VISUAL EMOTION ANALYSIS

Affective computing is an interdisciplinary field focused on enabling computers to recognize, interpret, and model human emotions by analyzing multimodal data such as text, speech, physiological signals (e.g., electroencephalogram, electrocardiogram), and visual stimuli [4], [53]. Its applications cover diverse domains, including human-computer interaction, mental health [41], and education [93]. Within this broader scope, visual emotion analysis (VEA) emerges as a specialized field concerned with emotional responses expressed by images.

Other studies often refer to related terms like affective image content analysis (AICA) [102], [101]. AICA explores the relationship between image content, emotions, and interconnected disciplines (e.g., psychology, sociology), as well as real-world applications. However, VEA [89] focuses narrowly on computational methods, which are feature extraction, representation learning, and classification, to analyze emotions evoked by visual stimuli - images, emphasizing algorithmic techniques for emotion prediction rather than interdisciplinary implications.

While AICA examines how images relate to emotions or affective states in disciplines, VEA prioritizes technical approaches, such as feature extraction, machine learning models, and dataset designs, to quantify emotional responses to visual data. This dissertation limits the scope of affective computing to visual emotion analysis, specifically addressing challenges in emotion recognition from static visual images for emotion classification.

Figure 2.1 displays stages that represent key components of VEA. Each stage will be described in the following sections.

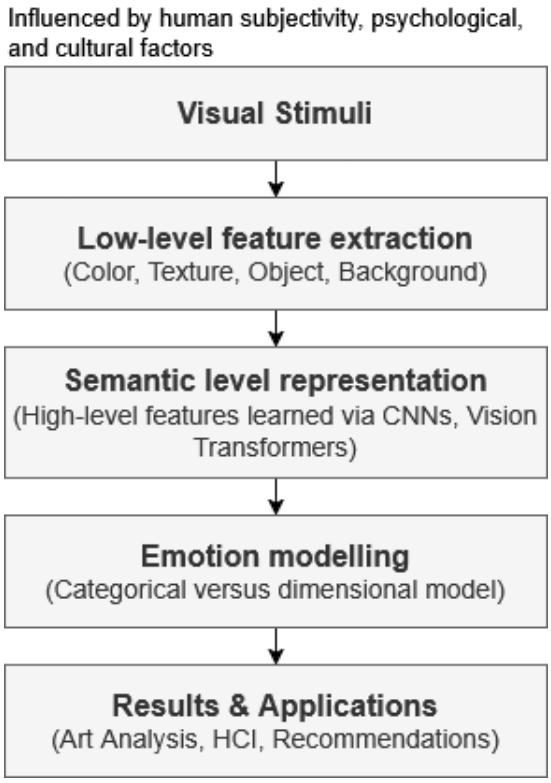


Figure 2.1: Each stage represents a key component of the VEA process. From feature extraction to emotion classification – highlighting the data flow and interdependency in emotion analysis pipelines.

Emotion recognition, also known as affective analysis, has been studied in various related forms, including text, speech, linguistics, music, sound, facial emotion expression, video, physiological signals, and multimodal data. VEA focuses on analyzing emotional content in images by extracting and interpreting visual attributes. The problem of emotion recognition can be divided into three steps: human annotation, visual feature extraction, and learning, where the extracted features are mapped to the perceived emotion. In various surveys it is said that the most significant difference between AICA and typical computer vision tasks is the affective gap concept. The affective gap can be illustrated by the following example: a physical rose object in a bright background might evoke a positive emotion, whereas in a dark, gloomy background, the expressed emotion is likely to be negative.

In VEA, emotions are typically represented using two types of taxonomies: categorical emotion states (CES) and dimensional emotion spaces (DES). These taxonomies define the emotion labels or dimensions used for deep learning, and determine how models are trained and evaluated. Most publicly available visual emotion datasets adopt one of these taxonomy type structures, which are foundational to both dataset design and model development.

Table 2.1: Models and their corresponding emotion states/dimensions [102].

Model	**Type**	**Emotion states/dimensions**
Ekman [20]	CES	Happiness, sadness, anger, disgust, fear, surprise
Mikels [49]	CES	Amusement, anger, awe, contentment, disgust, excitement, fear, sadness
Plutchik [58]	CES	(x 3 scales) anger, anticipation, disgust, joy, sadness, surprise, fear, trust
Parrott [57]	CES	A tree hierarchical grouping with primary, secondary and tertiary emotion categories
Sentiment [99]	CES	Positive, negative, (and neutral)
VA(D) [102]	DES	Valence-arousal(-dominance)
ATW [53]	DES	Activity-temperature-weight

Table 2.1 summarizes the most commonly used emotion representation models in visual emotion analysis. CES models provide a discrete set of labels, making them suitable for dataset construction and deep learning-based models. In contrast, dimensional emotion space (DES) models describe emotion continuously in terms of valence, arousal, and sometimes dominance, providing a more fine-grained representation of emotional states. Among DES formulations, the VAD model is the most widely used.

According to the authors in [37], seven basic groups of emotions are distinguished (joy, sadness, surprise, disgust, anger, fear, and neutral). In literature, facial emotion recognition consists of the following stages: image or photo pre-processing, face detection, facial feature extraction,

and facial expression classification [6], [16], [34], [63]. Emotions can be defined based on dimensions – their distribution in a certain two-dimensional or three-dimensional space. Emotions are described along several orthogonal axes, i.e., defining where they are located in two, three, or higher dimensional space. Dimensions ensure a unique identification of emotion and enable the expression of various emotions or groups of emotions. Two main dimensions of emotion are identified: arousal (physiological activation) and valence (emotional pleasantness). Arousal can be high or low, and valence can be positive or negative. Another dimension is also being described, called dominance. Dominance represents the degree of control ranging from controlled to in control [102]; however, it is rarely used due to its difficulty to measure. Furthermore, the dimensional model of emotion is a suitable representation of emotions and enables the assessment of similarity between emotional states [37], [19]. The primary drawbacks of DES are the difficulty in interpreting and distinguishing between continuous values of 2D or 3D models.

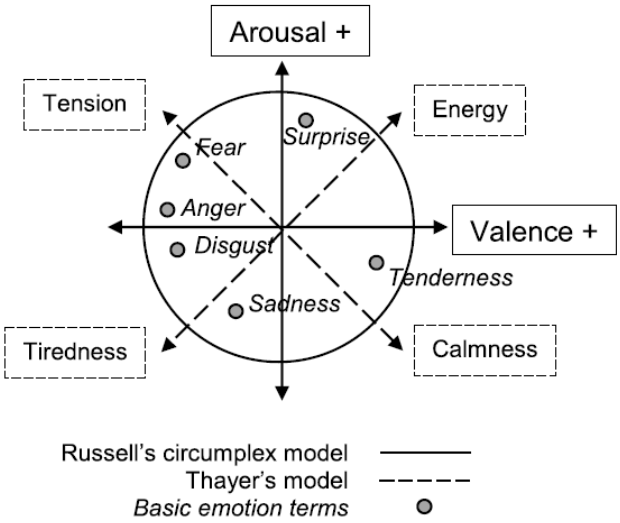


Figure 2.2: Two-dimensional model laying out the basic groups of emotions [19], [37].

Figure 2.2 displays a two-dimensional representation of emotions based on the valence-arousal model originally described by Russell

[67] and later refined by Thayer [76]. In this framework, valence refers to the pleasantness or unpleasantness of an emotion, while arousal describes the level of physiological activation. Emotions positioned near each other in this space share similar affective characteristics, whereas emotions located farther apart differ in both intensity and valence. This representation is widely used because it captures relationships between emotions and allows for continuous modeling of affective states.

Zhao et al. [99] emphasize that emotion recognition in general-purpose images operates at a higher level of abstraction than facial expression analysis. They highlight that emotions in images arise from global visual properties such as color, texture, shape, and compositional structure and are influenced by contextual and cultural factors. The notion of emotional semantics is introduced to describe how viewers interpret images based on cognitive models, cultural background, and aesthetic preferences. However, this process is highly subjective, making annotation difficult and leading to ambiguity in emotion labels. Zhao et al. also note that emotion responses to images have practical implications, for example, in marketing or multimedia analysis, where emotional cues may affect user decisions or engagement.

Plutchik's emotional model [58] illustrates how emotions can be organized within a multidimensional structure. The model conceptualizes emotions according to their similarity, intensity, and opposing relationships, providing a useful framework for understanding how different affective states relate to one another. In this formulation, emotions are often described along dimensions such as pleasure (valence), arousal, and dominance. The pleasure dimension distinguishes between pleasant and unpleasant emotions, such as joy and sadness, respectively. Arousal captures the level of activation or intensity associated with an emotion, ranging from low-arousal states such as calmness to high-arousal states such as ecstasy. Emotions with similar affective characteristics are positioned closer together, while opposing emotions are conceptually separated [99], [37].

Figure 2.3 illustrates a model of interpreting eight standard emotions. It is worth mentioning that emotions are typically first assessed on a sentiment basis. In the case of our further research, we decided to limit

Figure 2.3: Interpretation slice of emotions by Mikel [49].

ourselves to recognizing the visual image emotions constructed from Mikels' wheel of emotions.

It is worth noting that the subjectivity of people 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. 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.

2.2 APPLICATIONS OF VISUAL EMOTION ANALYSIS

Affective computing analyzes emotions or affective states from various modalities, including speech, physiological signals, images, and text. VEA is limited to the image modality. It is worth describing the affective computing potential application areas. Affective computing has significant applications in various areas, including healthcare, marketing, education, and opinion mining. Due to advances in affective image content analysis, it is gaining further attention. According to Resham et al. [4], affective computing is closely related to these research domains

such as physiology, psychology, linguistics, sociology, computer science, and mathematics.

With the advancements of smart devices, such as wearables and smartwatches, new healthcare-related applications emerge in the context of affective computing. There are various cases where emotion and sentiment recognition have a significant impact, such as mental well-being (stress, anxiety, and workload detection) [29], [35]. Recognizing emotional states using smart wearables can help users better understand their condition.

With the help of physiological signals or text-based surveys, it is possible to gather consumer preferences. This analysis could help better predict potential supply and demand chains. Furthermore, there is an interest in analyzing the impact of TV commercials on viewers. This may enable advertisers to gain insights into the effective emotional state of viewers when watching TV commercials.

Several studies have been done for the education application. During the learning process, physiological signals can be analyzed to assess the student's emotional state. This information could enable teachers to create a more effective teaching environment and gain a deeper understanding of their students. Another, easier approach is to analyze the facial expressions of students achieving similar aims. Furthermore, facial expressions can provide insights about student engagement [17].

These applications primarily rely on gathering and analyzing the physiological signals to predict emotional states or affect. However, there are also studies that explore application cases utilizing the visual image modality. A psychological image-based system is used for personality diagnosis for mental health. Visual and textual combined classification is used for analyzing the sentiment of product reviews [90]. From social networks, it is possible to explore the overall emotional experience, sentiment from uploaded photos. Analyzing photos from trips, emotional feelings can be related to physical surroundings [56].

Affective computing faces significant challenges in real world applicability. First, measuring physiological signals often requires specialized equipment and controlled environments, limiting scalability. Physiolo-

gical responses are ambiguous: similar signals (e.g., increased heart rate) can correspond to distinct emotions (e.g., excitement or anxiety), requiring multimodal approaches for reliable detection. Valence is difficult to capture using traditional sensors [4]. These limitations hinder practical deployment in real-world scenarios.

Affective computing faces notable challenges in real-world scenarios. Measuring physiological signals often requires specialized equipment and controlled environment, which limits applicability and scalability. Physiological measured responses are ambiguous: similar signals like increased heart rate can represent distinct emotions - excitement or anxiety. This aspect requires multimodal approaches, such as self-assessments, further increasing emotional ambiguity. Another dimension valence is difficult to reliably measure using sensors or detectors, because it aims to measure internal person's feeling.

Another challenge in affective computing is from cultural aspect. The same color can have different associations between different cultures, being associated with different concepts or norms [4], [12]. The same image can be perceived, understood and annotated differently across populations, increasing label ambiguity and reducing model consistency. These challenges are summarized in Table 2.2.

Table 2.2: Summary of key challenges in affective computing.

Challenge	**Description**
Cultural	Differences influence color associations and emotional perception [4], [12].
Dimensional	Valence (positive or negative) is difficult to measure [4].
Affective gap	Similar low-level features (e.g., pixel color) may not reflect the same emotion [99].

2.3 VISUAL EMOTION RECOGNITION METHODS

One of the primary goals of visual emotion analysis is to recognize the emotion expressed by a viewer in a given image. Other aims include feature analysis, object analysis of expressed emotion in images, and

practical adoption of visual emotion analysis [102]. This section aims to review methods, models, and approaches for the visual emotion recognition problem. The following study shows that facial expressions are not always the dominant cue in visual emotion recognition [70].

2.3.1 Visual Sentiment and Ontology-Based Methods

In the earliest work on visual emotion recognition, researchers relied on handcrafted, low-level visual descriptors that were classified using support vector machines (SVMs) [36]. Borth et al. later advanced the field by proposing the visual sentiment ontology (VSO) [7]. Based on Plutchik's wheel of emotions [58], the VSO was built by automatically mining user-generated tags on Flickr to extract a vocabulary of roughly 3,000 adjective-noun pairs (ANPs) - phrases that combine a sentiment-bearing adjective with a visually concrete noun. Each Flickr image tagged with a given ANP provides a weak (noisy) label, allowing the authors to train SentiBank, a set of linear SVM detectors. During prediction, the set of ANP detector scores can be mapped onto higher-level emotion categories, thus resulting in an evaluation of the overall visual sentiment conveyed by the image. DeepSentiBank was proposed to expand upon the previously described VSO model [10]. DeepSentiBank expands VSO by classifying ANPs using a convolutional neural network (CNN), improving performance over the SVM based SentiBank. The use of CNN significantly improves visual sentiment performance compared to the previous SentiBank model by Borth et al. [7]. However, these ontology-based models display weak generalization capability, necessitating the shift toward deep CNN models. However, the relationship between hand-crafted features and emotion is hard to establish. For example, it is a challenging task to model complex features, intra-class diversity, and high inter-class similarity [71].

2.3.2 CNN-Based Emotion Recognition

Current visual emotion analysis uses deep learning methods. The authors in [54] present their emotion classification results to recognize

emotions on people's faces. In their work, seven emotion classes are distinguished: neutral, happy, surprised, afraid, angry, sad, and disgusted. In the mentioned study, the authors perform emotion classification in a challenge called EmotiW 2015 [18] (Emotion in the Wild), comparing their results with those from a library. The authors employed a method known as transfer learning. Two data sets are used: FER-2013, which includes 30,000 grayscale photos (48x48 pixel resolution), and EmotiW, with images of 256x256 pixel resolution.The training process is divided into two stages: first, adapting the pretrained model architecture to the FER-2013 dataset; second, fine-tuning the pretrained network on the target task [66], [9]. After completing these training stages and using transfer learning, the authors achieve an accuracy of 48.5% in the validation set and 55.6% in the test set compared with 35.96% and respectively 39.13%.

CNN-based models noticeably improved emotion recognition accuracy compared to earlier hand-crafted methods. However, CNNs focus on the high-level semantic features, which leads to the loss of color, texture, and style elements important for bridging the affective gap in visual emotion analysis.

2.3.3 Multi-Level Representation Nodels

Other researchers employ multi-modal data from multiple sources for emotion classification [98], [30], [78]. Data of various types enable better classification results, where the data can include not only visual information but also audio signals when actors are required to perform (be filmed) and utter specific sentences. In another research, thermal infrared facial images are examined, and it is suggested that such images are difficult for humans to fake or imitate [78]. In the paper [30], the authors use what are called big data, consisting of audio and visual (filmed) signals. Big data are categorized into three types: happy, pain, and neutral states. For the classification of audio signals, the authors employ a 2D convolutional neural network, whereas for images, they utilize a 3D convolutional neural network. Visual signals undergo preprocessing to reduce data volume by selecting significant frames and skipping some in a set order. For the final classifier evaluation, the

authors explored model combination (fusion) strategies, including the Bayesian sum rule, Extreme Learning Machines (ELM), maximum, and product. The authors evaluated and determined that the best result with their big data was achieved using ELM fusion with 99.9% accuracy.

Zhang et al. [94] present an end-to-end model for image emotion prediction. They state that using local region information allows for improving the recognition performance. Their proposed end-to-end consists of these parts: a first classification stream, an emotion intensity prediction stream, and a second classification stream which outputs final emotion recognition results. They take advantage of class activation maps (CAMs), which are used to generate pseudo intensity maps from the first classification stream to guide the proposed network for emotion intensity learning. The authors derive these saliency maps, i.e., CAMs, from the pretrained first classification subnetwork. They also describe that the value of the pseudo-emotion intensity map CAM means the degree to which an area represents an emotion. According to the same authors, Zhang et al. [94] emotion intensity maps provide discriminative local information, which is used to improve emotion recognition performance in their work.

Zhang et al. [95] proposed a multiscale emotion representation network that combines a weakly supervised affective-region detection module with a kernel-based graph attention network for hierarchical feature fusion. Unlike their earlier method, this model incorporates contextual and multiscale information from both local regions and global image features, resulting in improved accuracy on datasets such as CAER-S and FI-8. However, its reliance on pseudo-region labels and the limited diversity of the CAER-S dataset raise concerns about generalization and robustness.

Another research addresses emotion recognition by learning multi-level representations that combine deep semantic information with shallow visual details. The authors claim that CNNs tend to focus on high-level semantics while not effectively used low-level visual cues, which are often important for expressing emotions. To address this problem, they propose a multi-level hybrid model that learns and integrates deep semantics and shallow visual representations for sentiment classifica-

tion. In addition, this study shows that class imbalance would affect performance as the main category of the affective dataset will overwhelm training and degrade the deep networks. Therefore, a new loss function is introduced to optimize the deep affective model. There are two fairly similar articles that present the same model as the authors are the same. The main novelty arising from the mentioned article and differing in that way from the previous research is the new introduced loss function. This loss function supposedly tries to address class imbalance in emotion images, which is undoubtedly prevalent.

Other researchers analyze multi-layered network models to recognize and classify possible visual emotions [85]. 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 [10], which produces conceptual emotion expression, e.g., a small beetle, which is expressed as the disgust expression. These expressions are formed as graph embedding in a 2D space. For the visual stream emotion recognition model, they use ResNet50 [28] model architecture. The final fused model is the combination of these two different model architectures, and the visual emotion predictions are obtained as a result. A similar approach and study were done by Zhang et al. [96], where a multi-level representation model with side branches, named Gram matrices for shallow features, is proposed. The authors in [96] attempt to integrate feature maps from different layers by applying a Gram matrix for further sentiment analysis – i.e., for negative and positive emotion classification. Zhang et al. [96] introduced the idea of integrating Gram matrix representations. This approach was later utilized by the authors of [51] for recognizing sadness emotion.

2.3.4 Multimodal and Notable Models

Xu et al. [82] 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. [45] 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 [65] 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. Recent multi-modal methods highlight integrating textual, visual, and contextual information using attention mechanisms or graph-based architectures. These models achieve high benchmark performance but introduce significant complexity. Another drawback is the requirement of multi-modal training data, which is difficult to obtain.

Previous studies by other authors describe multistage models for addressing potential issues in visual emotion recognition and classification [85]. When the image is passed to the network, these authors demonstrate that combining the visual-semantic stream with the visual stream yields better classification results. The visual-semantic stream constructs possible visual representations of emotions (embeddings), which are then fused with a visual stream whose output is a probability distribution over emotions. The visual semantic model is based on the structure of the DeepSentiBank model [10], which derives potential conceptual expressions of emotion, such as a small beetle that represents a feeling of disgust. These expressions are presented as graph embeddings in two-dimensional space. The visual stream part employs the architecture of the ResNet50 model [28]. The final structure of the model consists of the product of the outputs of these two models. The result of the model is a prediction of a visual emotion.

A similar conceptual idea and model are explored by Zhang et al. [96], who claim that shallow visual features are important for emotion recognition. According to the authors, such layers in the convolutional

neural network are at a shallow level, and deeper layers have the most significant influence on emotion recognition. The shallow level, in other words, refers to the layer depth, which is closer to the input of the original image. Therefore, they propose addressing this problem by using shallow layers, combining them with the formulation of the Gram matrix, and integrating these modules for the classification of emotional visual expressions. The authors note this as a sentiment classification problem.

Several authors [99] state that for the problem of emotion recognition, color, texture, shape, and contours are key features that describe visual emotions in a given image. This statement potentially suggests that emotion recognition in general nature images is a different problem than the commonly perceived emotion recognition in facial expressions. Detected and recognized emotional expression may not have a physical origin because the features that define emotion might be broad and varied.

There are various convolutional neural network models designed for object detection and recognition. Due to their prominence and impact, the following CNNs are going to be considered: ResNet, Xception, and EfficientNetV2.

One of the most influential convolutional neural networks – ResNet – was introduced to address the vanishing gradient problem in very deep architectures through residual blocks using skip connections [28]. These residual connections enable stable training and have made ResNet a widely used backbone in VEA. More recent architectures, such as Xception and EfficientNetV2, improve convolutional efficiency but have seen limited adoption in VEA.

ResNet type networks are also widely used for visual emotion analysis. Researchers use ResNets as a pretrained model for feature extractors, integrate these extracted features for further enhancements in emotion recognition [45], [44], [96], [95], [81].

Another convolutional neural network used in our studies is called Xception [11]. The author of this network introduced an extension of standard convolution known as depth-wise separable convolution,

which achieved better results than the InceptionV3 [72] model. Depth-wise separable convolution refers to a spatial convolution performed independently across each channel, followed by a point-wise convolution – a 1×1 convolution filter operation. The improvement offered by the authors comes from a different sequence of operation actions. Depth-wise separable convolution also acts as a factor that reduces the impact of overfitting and accelerates the training process. The Xception network comprises three main components: the entry flow, the middle flow, and the exit flow. The Xception model has found some use for visual emotion analysis [86], [42]. However, the Xception network has not been widely adopted and has been outperformed by a newer model.

The next CNN is called EfficientNetV2, which was introduced in 2021. According to its creators, these family of convolutional neural networks proved to be the best in the ImageNet recognition library benchmark [38], [74]. The main improvements of EfficientNetV2 are highly optimized convolutional blocks, which were inspired from the MobiletNetV2 model [68]. Relatively few studies in the literature are related to the recognition of visual emotions in images using the EfficientNetV2 models. One such study investigates visual emotion states using an older version of the EfficientNet model [5]. Thus, there is an opportunity to investigate visual emotion recognition using EfficientNetV2.

2.3.5 Loss Functions and Representation Learning Advances

Another important research direction in visual emotion analysis concerns the design of specialized loss functions that better capture semantic relationships and perceptual distances between emotion classes, thereby improving class separability and recognition performance. Standard image classification tasks commonly rely on categorical cross-entropy (CE) loss for training [75]. Yang et al. propose an emotion-circle-based representation learning framework for VEA [87]. They introduce a progressive circular (PC) loss, which progressively restricts on three emotion attributes: polarity, type, and intensity. The PC loss penalizes differences between predicted and ground-truth emotion vectors in a coarse-to-fine manner, encouraging structured representation learning. The final training objective combines cross-entropy (CE), Kullback–Leibler (KL)

divergence [75], and the mentioned PC loss to utilize relational dependencies between emotions within an emotion circle structure. In a subsequent work, Yang et al. [88] decompose visual emotions into components, including color, object, and face cues, and employ a hierarchical cross-entropy loss that assigns different penalties to classification errors according to the hierarchical distance between emotion categories. This non-standard CE formulation penalizes misclassifications of semantically distant emotion classes more strongly than nearby ones. Xu et al. propose a semantic graph prompt learning module that functions as a high-level semantic filter, enhancing the emotional interpretation of visual features extracted at lower levels [83]. Their framework separately optimizes low-level visual features using frequency filtering and high-level semantic representations through an emotion-aware graph with prompt embeddings. The resulting representations are concatenated into a unified vector used to predict the emotion distribution. They further introduce a dual-loss optimization objective consisting of KL divergence and CE terms, where joint optimization encourages both full distribution alignment and consistency with the dominant emotion category. Bustos et al. investigate how to leverage CLIP's joint embedding space for visual sentiment analysis (VSA) [8]. Their approach adapts CLIP's large-scale multimodal representations for emotion recognition using two variants: CLIP-E Contrastive and CLIP-E cross-entropy. The CLIP-E Contrastive variant modifies CLIP's original contrastive loss [60] to align image embeddings with sentiment-oriented textual prompts. They introduce three types of textual prompts – sentiment captions, synonym-based phrases, and image descriptions to better capture the emotional semantics of visual content.

Sun et al. [71] propose a supervised contrastive learning-based model for classifying image emotions. Their model integrates low-level, handcrafted features (extracted using the LBP-U algorithm [55] – Local Binary Patterns) and deep emotional features (learned through a ResNet50 encoder [28]), 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 enhance feature discrimination by reducing intra-class variability and improving inter-class separability. In the second stage, the pretrained encoder is frozen, and

the classifier is trained using cross-entropy loss optimization. Experimental results demonstrate performance gains over baseline methods on a single image emotion dataset, highlighting the benefits of supervised contrastive learning. However, the study does not evaluate cross-dataset generalization or isolate the specific contribution of the contrastive optimization to the final performance.

2.4 CONCLUSIONS OF THE CHAPTER

The literature review shows that visual emotion recognition remains a challenging problem due to subjective human perception, cultural differences, and the affective gap between image features and emotional meaning. Although numerous psychological models and taxonomies exist, categorizing emotions in images remains challenging to define and annotate reliably. This highlights a need for methods that can handle ambiguity and variation in emotional interpretation.

The review also shows that deep learning has become the dominant approach in this field. Classical hand-crafted and ontology-based methods (VSO, ANPs) provide limited generalization, whereas CNN-based models capture richer semantic information. Recent studies further propose multiscale features, attention mechanisms, region detection, and contrastive learning. However, these methods still face challenges such as losing stylistic information, limited dataset diversity, and difficulties in separating visually similar emotion classes.

Overall, the literature review identifies existing limitations and opportunities for methodological improvement. It provides the basis for developing a model that combines semantic and stylistic features and improves class separability through enhanced optimization.

3 ANALYTICAL REVIEW AND COMPARISON OF CNN MODELS FOR VISUAL EMOTIONS CLASSIFICATION

This chapter reviews and compares widely used convolutional neural network architectures and the main methodological components required for visual emotion classification in general-purpose images. ResNet50, Xception, and EfficientNetV2 are compared. Any of them is suitable as a backbone. The goal is to select an appropriate backbone for developing a new model. The chapter presents the choice of backbone architecture and datasets used in the later methodology and experiments. Establishing these choices is necessary because both the backbone design and the datasets directly affect training stability, baseline performance, and the validity of comparisons when introducing architectural and optimization improvements. In this thesis, the topic of visual emotion recognition is addressed. We consider only categorical emotion states for the classification problem. In our case, we refer to emotion categories from Mikel's model [49]. The chapter is partially based on the results published in [A1, A2, B1].

3.1 DEEP LEARNING COMPONENTS AND BASELINE CONFIGURATION

Deep learning can be categorized into several broad components:

- Dataset: collection, preprocessing, augmentation [25].
- Model: architecture design, complexity, and improvements [39].
- Loss function: selection and design of a suitable objective function to minimize [75].
- Hyperparameters: selection of learning rate, batch size, effective training strategy [92].

In this chapter, these components are specified to define a consistent baseline configuration that will be used throughout the subsequent

methodology and ablation studies. These broad components are typically referred to as deep learning. In this dissertation, visual emotion analysis is also described using deep learning (DL). These fields have different impacts and roles in the DL landscape. Our main contributions arise from model improvements and the expansion of the training process.

3.2 RESNET MODEL OVERVIEW

One of the most popular and foundational models is the ResNet convolutional neural network [28]. It might be valuable to outline the key structure of the ResNet model.

Table 3.1: Structure of ResNet50. Each bottleneck block has a $1\times1 \rightarrow 3\times3 \rightarrow 1\times1$ conv sequence.

Stage	**Operation**	**Stride**	**Channels**	**# Blocks**
0	Conv7×7 + MaxPool3×3	2	64	1
1	Bottleneck k1-3-1	1	256	3
2	Bottleneck k1-3-1	2	512	4
3	Bottleneck k1-3-1	2	1024	6
4	Bottleneck k1-3-1	2	2048	3
5	AvgPool + FC	–	1000	1

Table 3.1 presents the overall structure of the ResNet-50 network. The input passes through six sequential stages. Stages 1-4 consist of multiple bottleneck blocks (e.g., Stage 2 contains four blocks), each implementing a $1\times1 \rightarrow 3\times3 \rightarrow 1\times1$ convolutional sequence. The sequence of 1×1 refers to the spatial dimensions of the kernel performing convolutions. The stride parameter controls spatial downsampling through the convolution kernel's step size. ResNet employs two downsampling approaches: pooling operations (used in Stages 0 and 5) and strided convolutions (with a stride greater than one in the bottleneck blocks of Stages 2-4). The ResNet model remains the default choice for transfer learning problems, due to its relatively simple yet efficient structure. ResNet models are still widely used in research and in practical applications [84], [22].

3.3 XCEPTION MODEL OVERVIEW

Another convolutional neural network used in our studies is called Xception [11]. The authors of this network introduced a variant of the standard convolution, known as depthwise separable convolution, which achieved better results than the InceptionV3 model [72]. Depthwise separable convolution refers to a spatial convolution performed independently across each channel, followed by a pointwise convolution - a 1×1 convolution filter operation. The improvement offered by the author comes from a different sequence of actions. This convolution also acts as a solution for reducing overfitting and accelerating the training process. The Xception network consists of three main components: the input passes through what is called an entry flow, then into a middle flow where the module is repeated eight times, and finally through an exit flow. Xception modules also utilize the possibility of residual connections, which, according to the author, provide a comparatively higher accuracy on the ImageNet validation set than when these additional connections are not used.

Figure 3.1 shows the structure of the middle flow of the Xception network. This flow can be understood as an arrangement of repeating, defined layers based on depthwise separable convolution layers. The input to this flow will be feature maps of $19 \times 19 \times 728$ dimensions, with the initial input images of $299 \times 299 \times 3$. The output of the depthwise separable convolution specifies a predefined number of channels. The output and input feature maps of this block are connected using a residual connection, also known as a skip connection.

3.4 EFFICIENTNETV2 MODEL OVERVIEW

EfficientNetV2 [74], one of the newer models in the convolutional neural network family was introduced in 2021. According to the authors of EfficientNetV2, this network has achieved the best results in the ImageNet classification challenge [66]. The ImageNet ILSVRC2012 dataset comprises 1,281,167 training images, 50,000 validation images, and 100,000 test images, aiming to classify 1,000 categories from the aforementioned

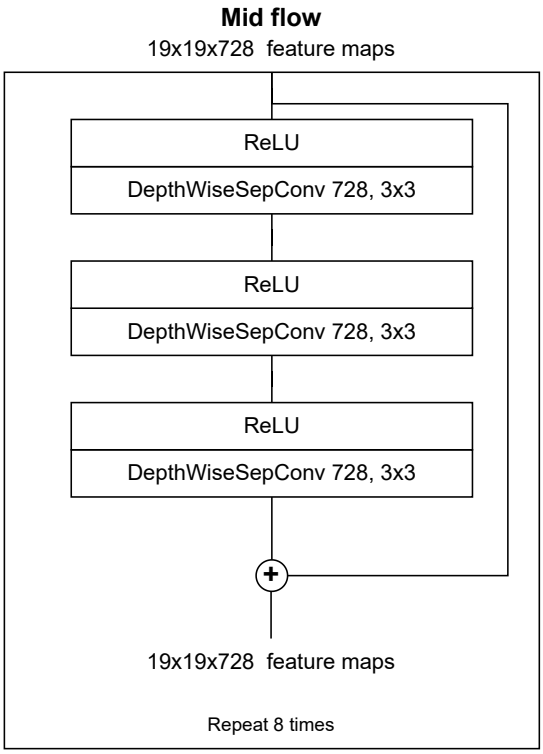


Figure 3.1: Structure of the middle flow of the Xception network. Source: [11].

set. EfficientNetV2 family models achieve better results than previous solutions because they incorporate more efficient blocks called MBConv and Fused-MBConv. The authors also conducted a neural network architecture search to find optimal network parameters, using their older EfficientNet [73] B4 version as a base, resulting in a model called EfficientNetV2S [74]. The search space method is often referred to as neural architecture search - selecting exemplary architectures from a search space, evaluated by a chosen metric (accuracy). In essence, this method optimizes the entire network structure and is therefore often referred to as global search optimization, which implies that the search space is very large and computationally expensive. Model optimization and selection (sampling) for the EfficientNetV2 module were based on these objectives: accuracy, training speed, and number of parameters. EfficientNetV2B0

and EfficientNetV2B2 are scaled-down versions of the original EfficientNetV2, featuring fewer parameters, fewer convolutional layers, and trained on lower-resolution images. The same authors also introduced progressive image resolution changing combined with adaptive regularization training methods, which significantly reduced the time required for training not only for their presented model but also for existing older models. The novelty and main idea of progressive training are to divide the training phase into several smaller steps – initially training the network using lower-resolution images with weaker regularization and, in later stages, increasing the image resolution and incorporating stronger regularization using mixing [97] (blending images into one and outputting a probabilistic category), random augmentation [13], and stochastic dropout [69].

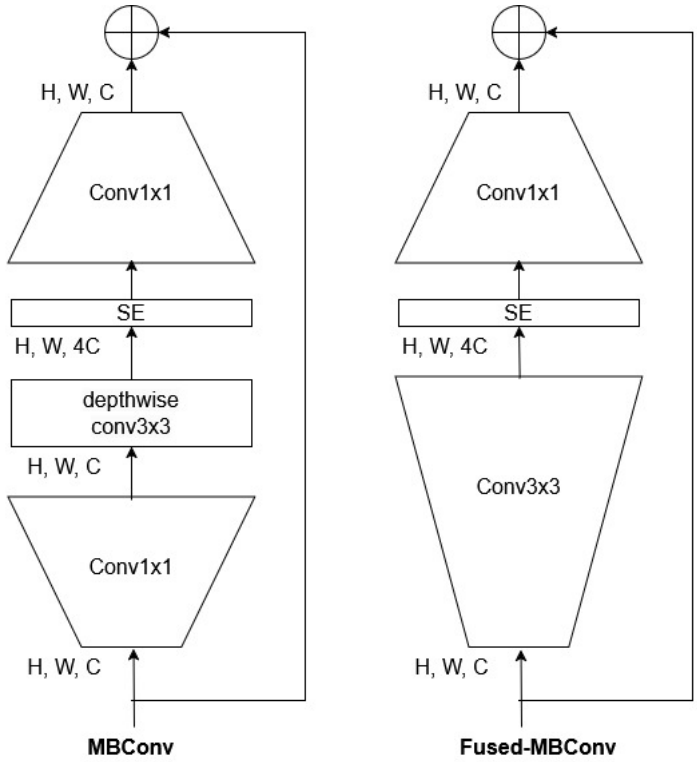


Figure 3.2: Structure of MBConv and Fused-MBConv blocks. Source: [74].

Figure 3.2 shows the structural blocks of the EfficientNetV2 architecture. By using efficient calculations and combinations from these

modules, the authors determined the entire network structure. The MBConv block, also known as the inverted residual block, is understood as a variation of the residual block designed to achieve higher efficiency. The inverted residual block was first introduced in the MobileNetV2 convolutional neural network architecture [68]. 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. This normalization is necessary to combine them using a residual connection (skip connection). The authors of EfficientNetV2 also enhanced this block with a so-called squeeze-and-excitation (SE) layer, which was first introduced by the authors of [31]. This layer, essentially a module, consists of a global average pooling operation, a fully connected layer with ReLU activation, a subsequent fully connected layer with sigmoid activation, and multiplication operations. According to the authors of the original presentation, this block helps achieve better results in benchmark solutions with a minimal increase in computational cost. The essential difference between the modules is that Fused-MBConv replaces the first two layers with a conventional 3x3 convolution. For simplicity, the full structure of the blocks is not shown. It is intended to show only those layers or operations that are significant in terms of computation.

Table 3.2: Structure and parameters of EfficientNetV2S. MBConv and Fused-MBConv blocks are described in Figure 3.2. Diagram source: EfficientNetV2: Smaller Models and Faster Training [74].

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	Convlxl & Pooling & FC	-	1280	1

Table 3.2 illustrates the structure of EfficientNetV2S detailing its com-

ponents and blocks. The structure was optimized using reinforcement learning based on the ImageNet dataset [66]. Stride refers to the step size of the convolution operation. Channels No. indicates the number of output channels from a particular block or operation. Layers No. specifies the count of particular block repetitions within a certain stage. For example, the number of layers in the fourth stage – six – indicates the number of repetitions of the MBConv block. MBConv[n] denotes the block MBConv with an expansion factor of n - the initial 1×1 convolution receives C channels and expands the output to nC channels. Here, $k3 \times 3$ refers to the spatial dimensions of convolutional filter (kernel). SE0.25 refers to the reduction ratio of the squeeze and excitation block used to model channel-specific relations.

The ResNet50, Xception, and EfficientNetV2 models are designed for similar computer vision problems. They achieved very good classification results on the ImageNet-1k dataset [66]. The main question regarding the mentioned CNN models is choosing one for further visual emotion analysis. The following sections describe the datasets and the baseline experimental setup used to select a backbone.

3.5 DATASETS

The next part discusses suitable visual emotion datasets for classification.

Several datasets are introduced for visual emotion analysis and exist. They include ArtPhoto, IAPS-subset, and abstract paintings [46], [100]. In total, these datasets provide approximately 1,400 visual emotion images. Due to a lack of sufficient images to train deep neural networks, the Flickr and Instagram dataset (FI-8) was constructed. As modern emotion recognition models require substantially larger and more diverse training data, larger-scale datasets were introduced.

The WEBEmo dataset was formed with the goal of avoiding the inherent bias in datasets by collecting a large number of general-purpose emotional images [61]. This dataset is divided into three hierarchical groups: starting with the first level, which includes positive and negative, the second level comprises eight emotions of happiness, anger,

surprise, satisfaction, disgust, excitement, fear, and sadness, and on the third level, there are 25 detailed categories. The researchers then generated a list of significant words to obtain images from publicly available online sources. This approach resulted in a dataset of approximately 256,000 general-purpose emotional images. We successfully downloaded and acquired 220,854 images from that dataset.

Table 3.3: Statistics of downloaded images from the WEBEmo dataset.

Positive Emotions	**Images**	**Negative Emotions**	**Images**
Happiness	50,194	Anger	25,227
Surprise	25,094	Disgust	17,314
Satisfaction	35,607	Fear	14,620
Excitement	26,954	Sadness	25,847
Total Positive	**137,847**	**Total Negative**	**83,007**

Table 3.3 presents the statistics of the visual emotion images downloaded from WEBEmo. It was decided to use this dataset for further research due to the large variety and volume of general-purpose emotion images.

To focus on the recognition of sadness and to reduce the impact of class imbalance, the WEBEmo dataset was used to construct a smaller binary dataset consisting of images expressing sadness and images representing all other emotions.

Figure 3.3 presents an example of the constructed smaller dataset from WEBEmo. Two groups of visual emotions are distinguished: sadness and others. The other classes include emotions that may differ from the visual category of sadness. It is worth noting that in these examples (and in all others), the human face is not the primary subject of analysis in the image. It is also worth noticing that in at least three out of the 16 images shown, there is some text in the image, which itself reflects meaningful information. This textual information is not relevant when investigating visual emotions in general-purpose images. The next section discusses the decision-making process for selecting and discarding images that contain any text.

Figure 3.3: Example images from the smaller dataset constructed from the WEBEmo dataset.

3.5.1 Discarding Textual Information From the Visual Image Datasets

Upon discovering that the WEBEmo dataset contains a significant number of images with corresponding text, it was decided to construct a model that filters out images with text. The idea behind this model is to determine whether text is present in an image. A binary classification model was employed to address this issue.

The dataset was divided into two distinct groups: images containing any relevant text and images without text. The dataset used to identify text in images was sourced from the work of Gomez et al. [24]. From this dataset, it was possible to download 6,627 images. The text in this dataset appears in various forms and expressions, including billboards, facades, shop windows, and book covers, among other formats.

For the dataset where text is not present in images, the dataset from Roy et al. [64] was used. From this dataset, it was possible to download 9,720 general-purpose images without any text. Essentially, there exists a wide variety of choices for selecting images for this category because there are naturally more images without text.

Xception CNN was chosen to solve a binary task: detect whether an input image contains any text. Training relied on a custom two-class dataset ("text" versus "others"). The overall dataset is divided into three sets - a training set that represents 70% of the total, a validation set making up 15%, and a test set comprising the remaining 15%. The training set is used solely for network training, the validation set is intended to assess the suitability of model parameter selection, and the test set is used to evaluate the model's effectiveness with unseen data. The images in the training set were further augmented to prevent the models from overfitting to the training images. Random zoom was applied to these images with the following values: a height factor, meaning vertical zoom, between -0.05 and -0.1, which in percentage terms represents zooming in at a random value in the range [+5%, +15%]. The width factor signifies a random horizontal zoom with the same values chosen. Then, a random rotation was selected with interval values ranging from $[-0.6\pi, 0.6\pi]$. A random flip in either the horizontal or vertical direction was also used.

Pre-training using the ImageNet dataset [38] allows for achieving better accuracy faster and more efficiently than when training a network from scratch. For optimizing network training, the stochastic gradient descent (SGD) was chosen, and the learning rate was set to 0.001 following several preliminary trials. The loss function used was sparse categorical cross-entropy. We used sparse categorical cross-entropy because the dataset class labels are integer-encoded (from 0 to $C-1$), which eliminates the need for one-hot label encoding.

Figure 3.4 presents the confusion matrix of the trained text recognition model. An F1 metric score of 99.9% was obtained with the labeled testing data. This score essentially is a limited indicator of whether the model will be able to detect textual information from images of different origins.

Figure 3.5 shows examples of images that the trained text recognition model identified as containing text. These images were taken from the smaller WEBEmo dataset constructed as mentioned in the previous section. Text is not always detected correctly in all images. In the large-scale WEBEmo dataset, we lack prior information to determine which

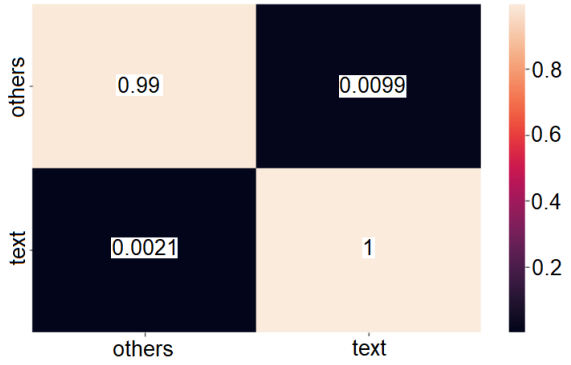


Figure 3.4: Confusion matrix of the trained text presence classification model.

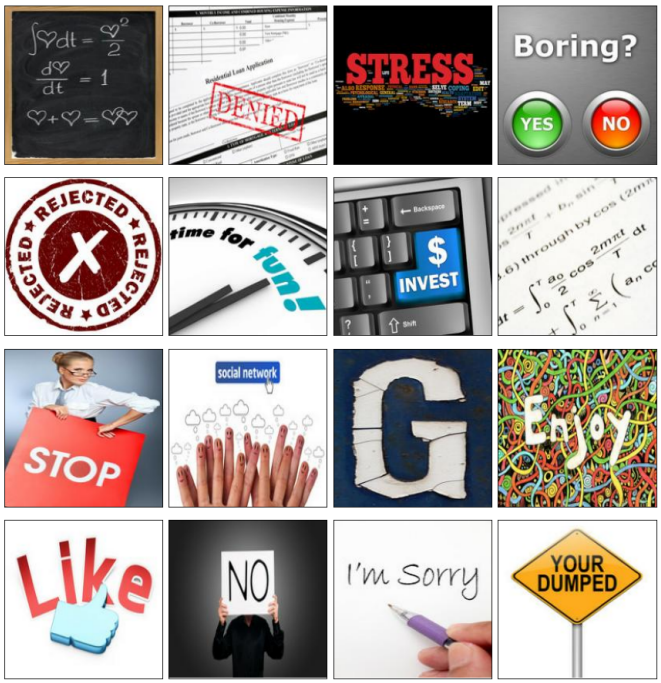


Figure 3.5: Examples of images identified by the model as containing text. Images are taken from the WEBEmo dataset.

images contain text or not; therefore, an automatic solution has been applied. From the originally downloaded WEBEmo dataset of 220,854 images, we automatically obtained a filtered dataset of 186,777 images containing no text.

Finally, after the additional undersampling, we have obtained a 61,074 filtered images dataset, where about 46% of images represent sadness emotion. The constructed dataset has been divided into 80% training, 10% validation, and 10% testing splits. The WEBEmo training subset contains 26,445 images expressing no sadness emotion, and 22,413 images conveying sadness emotion. Similarly, the WEBEmo validation subset consists of 3,284 images expressing no sadness emotion, and 2,823 images expressing sadness emotion. Finally, the testing subset split is divided into the same ratio as validation subset, consisting of the same number of images in each class. From now on this dataset should be called WEBEmo sadness, emphasizing that dataset has been constructed from the WEBEmo [61] visual emotion set.

The described WEBEmo sadness dataset alone is insufficient to address visual emotion recognition challenges.

3.5.2 General-Purpose Visual Emotion Datasets

To assess the network's generalization capabilities, the Flickr and Instagram (FI-8) dataset [91] was also used. It comprises 23,308 images labeled according to Mikels' emotion hierarchy [49]. The dataset was constructed by gathering potentially suitable images from Flickr and Instagram using the eight emotion categories of Mikels' model, after which each image was labeled by five hired Amazon Mechanical Turk workers. Images receiving more than three votes for a label were kept in the final dataset.

Further, we evaluated the network on EmoSet [89], a large-scale visual emotion dataset designed for 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 118,102 images, where each image is labeled according to one of eight emotion categories based on Mikels' model [49].

The emotion category distribution is given in Table 3.4.

Table 3.4: Emotion distribution in EmoSet-118K and FI-8 datasets.

Positive	Amusement	Awe	Contentment	Excitement
EmoSet-118K	19,445	15,037	16,337	19,828
FI-8	4,924	3,151	5,374	2,963
Negative	Anger	Disgust	Fear	Sadness
EmoSet-118K	10,660	10,666	13,453	12,676
FI-8	1,266	1,658	1,032	2,922
Dataset	Total			
EmoSet-118K	118,102			
FI-8	23,308			

This shows a relatively even distribution across both positive (amusement, awe, contentment, excitement) and negative (anger, disgust, fear, sadness) emotions, which makes EmoSet-118K suitable for further evaluation. The EmoSet-118K dataset was annotated by human annotators, with 10 annotators assigned to each image. A final ground-truth label was assigned when more than seven out of the ten annotators agreed on the same label. The most frequent categories are amusement and contentment, whereas the least frequent are anger and disgust.

We also employed the CAER-S (Context-Aware Emotion Recognition - Static) dataset [40], a subset of the larger CAER dataset. CAER-S comprises training and testing sets without a validation subset and includes the following emotion categories: anger, disgust, fear, happy, neutral, sad, and surprise. Notably, the training set contains exactly 7,001 images per category, and the testing set contains 2,999 images per category, indicating a perfectly balanced distribution. Figure 3.6 displays sample images from this dataset. The authors of CAER-S [40] indicate that the full dataset was obtained from emotion-related scenes

Figure 3.6: Visual emotion examples in the CAER-S dataset.

collected from 79 TV shows. CAER-S was created by extracting static images from this larger annotated dataset.

Table 3.5: Summary of visual emotion datasets used in the dissertation.

Dataset	Classes	Total images
WEBEmo sadness	2	61,074
FI-8	8	23,308
EmoSet-118K	8	118,102
CAER-S	7	69,999

Table 3.5 shows a summary of the visual emotion datasets used in this dissertation. FI-8 is the smallest of the four datasets. However, at the time of its release in 2016, FI-8 was among the largest publicly available labeled visual emotion datasets. More recently, EmoSet-118K, introduced in 2023, represents a substantial increase in the scale of labeled emotion images [89].

3.6 IMPLEMENTATION DETAILS

The remaining question is to choose a suitable CNN for visual emotion analysis. The selection of these three previously described CNNs needs to serve such functions:

- Be a baseline network for enhancing visual emotion classification.
- Act as a foundational block for designing an effective visual emotion recognition model.

Training models and evaluating their performance metrics are key steps in selecting a CNN.

3.7 EVALUATION OF MODEL PERFORMANCES

One of the key concepts behind deep learning is defining a loss function. During training, the loss function is a measure of the difference between the predicted outputs of the model and the expected (true) outputs [75]. Training seeks to minimize that difference by optimizing model parameters. We use an objective loss function for CNN training is called sparse categorical cross-entropy loss [25]. The sparse categorical cross-entropy loss is appropriate for emotion classification problems with integer-valued class labels. Let C denote the number of visual emotion classes and let $\mathbf{y} = (y_1, \ldots, y_N)$ be the vector of ground-truth labels for a dataset of N samples, where $y_k \in \{1, \ldots, C\}$. The loss is defined as

$$L_{\text{entr}} = -\frac{1}{N} \sum_{k=1}^{N} \log(p_{k,y_k}), \tag{3.1}$$

where p_{k,y_k} denotes the predicted probability assigned to the ground-truth class y_k of sample k.

3.8 DATA AUGMENTATION AND PREPROCESSING

Another key concept behind deep learning is the use of hyperparameters for model training. This section describes data augmentations and preprocessing steps for training described CNNs.

The dataset-dependent hyperparameters are the learning rate and the number of training epochs. We train with stochastic gradient descent (SGD) using a momentum of 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 [43]: it decays from 0.02 down to 1×10^{-4} and restarts back every five epochs. During training, first apply a random resized crop to 224×224 and a random horizontal flip. Then, the RandAugment augmentation technique [13] with a distortion strength of three and a number of transformations of two was used.

3.9 EXPERIMENTAL RESULTS

The experiments were carried out using PyTorch deep learning library.

Table 3.6: Accuracy results with standard deviations for different CNNs on four emotion datasets. The results were averaged over three training runs and evaluated on the corresponding test sets.

Dataset	ResNet50	Xception	EfficientNetV2S
WEBEmo sadness	$81.34 \pm 0.21\,\%$	$81.31 \pm 0.15\,\%$	$81.33 \pm 0.22\,\%$
FI-8	$68.45 \pm 0.80\,\%$	$70.03 \pm 0.21\,\%$	$68.06 \pm 0.29\,\%$
EmoSet-118K	$77.43 \pm 0.33\,\%$	$77.55 \pm 0.05\,\%$	$77.67 \pm 0.08\,\%$
CAER-S	$82.98 \pm 0.25\,\%$	$88.15 \pm 0.06\,\%$	$90.40 \pm 0.17\,\%$

An experimental study evaluating CNNs on various visual emotion datasets was conducted, and the results are presented in Table 3.6. The three models show fairly similar performance on the first three datasets, except for the CAER-S. The most notable differences are observed

in the accuracy results for the CAER-S dataset. The EfficientNetV2S neural network achieved the highest accuracy by a clear margin on the CAER-S dataset. For practical reasons, EfficientNetV2S is selected as the backbone because it provides competitive performance across datasets while remaining computationally efficient and flexible to extend. Its architecture also supports the use of a single consistent backbone for the ablation studies and model improvements introduced in the next chapter.

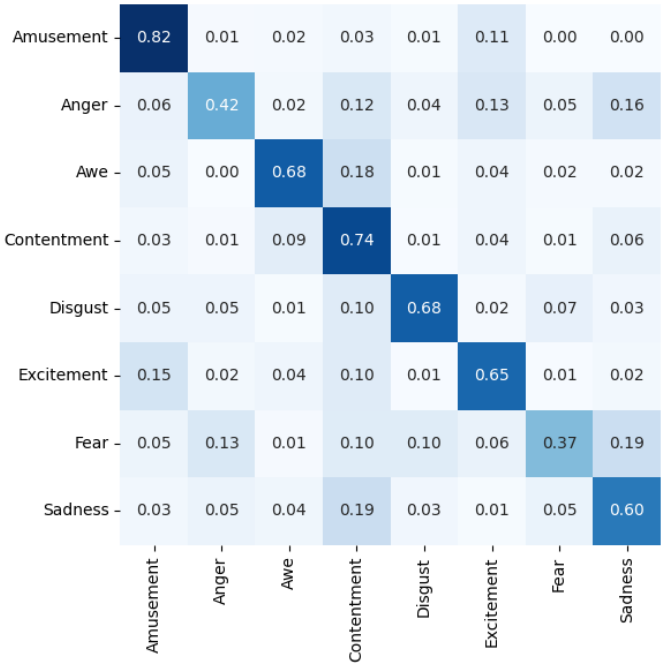


Figure 3.7: EfficientNetV2S trained model results on FI-8 testing set.

The trained CNNS were additionally evaluated using confusion matrices. In Figure 3.7, the weakest recognition performance is observed for the fear and anger visual emotion classes. Furthermore, the confusion matrix shows that the trained EfficientNetV2 network favors the contentment class, which is the majority class in the FI-8 dataset, The contentment emotion is visually similar to awe, and the trained network struggles to discern these emotions. Similarly, the same applies to fear and sadness emotions, whose visual differences are fairly subtle.

Figure 3.8 presents the results of the trained model. The weakest

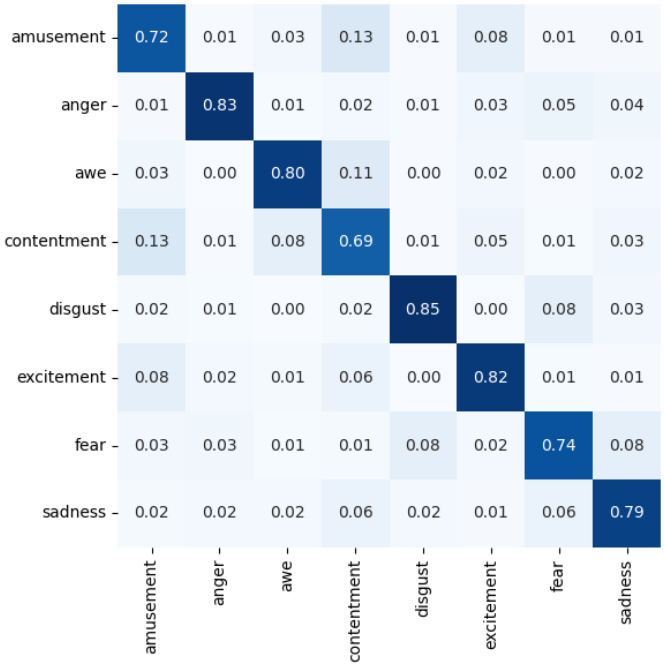


Figure 3.8: EfficientNetV2S trained model results on EmoSet-118K testing set.

performance is observed for the contentment and amusement emotion classes. A notable difference is the stronger recognition performance for fear and sadness emotions, which is a distinct improvement over the previous case. This might indicate that the EmoSet-118K dataset is of better quality. Class distributions are fairly balanced, and emotions are well represented. The contentment class still remains challenging to recognize because of its visually similar features to other positive emotions.

After constructing the potential dataset to solve the text detection problem, the EfficientNetV2 [74] model variant was chosen for network training and utilized as a feature extractor for the proposed methodology, serving as the backbone of the developed model.

3.10 CONCLUSIONS OF THE CHAPTER

This chapter reviewed widely used convolutional neural network architectures and the main methodological components required for visual emotion classification in general-purpose images. By comparing ResNet50, Xception, and EfficientNetV2 and describing the supporting pipeline (emotion datasets, text filtering, and augmentation strategies), the chapter established the practical criteria for baseline model selection, performance consistency across datasets, and architectural efficiency. The analysis also highlighted typical limitations of baseline CNNs for emotion recognition, including class confusion among visually similar emotions. EfficientNetV2S provided a better performance across datasets while remaining computationally efficient and flexible to extend. However, any other of the compared CNNs can be used as a backbone, too.

The results of this chapter lead to the development of a new CNN-based model presented and investigated in the next two chapters. We develop the model so that it would be independent of a particular backbone architecture and can be integrated with any modern convolutional neural network pretrained on large-scale image datasets. The backbone is not the primary focus of this work. Therefore, the analysis of possible backbones, including experiments, is performed in Chapter 3, which precedes the presentation and experimental investigation of a new model in Chapters 4 and 5. Backbone serves as a feature extractor whose representations are subsequently refined through the proposed training strategy. The investigation in this chapter led to choosing EfficientNetV2S as the backbone. In the next chapters, this backbone is used as the feature extractor in the proposed CNN-based model, with additional Gram matrix modules and training strategy introduced to improve the recognition of visual emotions.

4 NEW MODEL FOR VISUAL EMOTION RECOGNITION

In this chapter, a methodological framework is presented for the application of Gram matrix modules to improve visual emotion analysis. This problem is formulated as a visual emotion classification task. Additionally, this chapter proposes model improvements designed to enhance the robustness of emotion recognition performance. Visualization techniques are used to explore and analyze the behavior of the model. This allows a qualitative assessment of the relationships among learned features and gaps within the developed visual emotion recognition model.

The research framework methodology and its components have been published in peer-reviewed papers [A1, A2, B1].

4.1 PROPOSED MODEL AND TRAINING STRATEGY

This section introduces the proposed model and training strategy for improving visual emotion recognition performance. Building on the findings of the literature review (Chapter 2), which highlighted the difficulty of capturing semantic and low-level visual information, this chapter presents the model design and the integration of contrastive-center loss into training.

The proposed CNN-based model in this dissertation builds upon prior work by Zhang et al. [96], who demonstrated that shallow-layer Gram matrices effectively capture texture and color features for binary sentiment classification. In addition, we extend the research from binary sentiment classification to multi-class visual emotion classification. We further propose more compact Gram matrix modules that extract feature representations at multiple depths of a CNN.

Additionally, visualization techniques are employed to analyze model behavior and qualitatively assess feature representations across emotional categories.

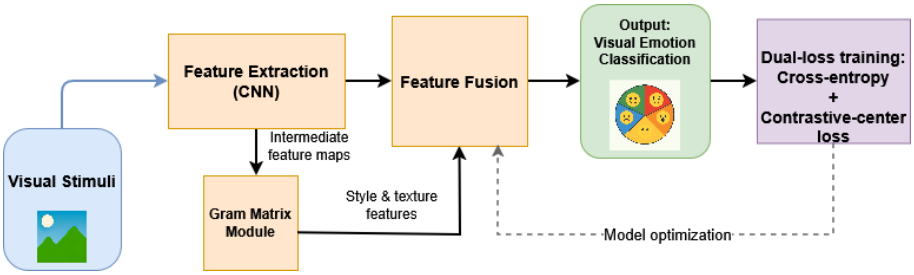


Figure 4.1: Proposed CNN-based model for VER.

We propose auxiliary expandable Gram matrix modules, which extend from the main CNN, and we combine the gathered features into the final network. Our main contributions and improvements, compared to the previous work of Zhang et al. [96], are as follows:

- Develop a CNN-based model that incorporates expandable Gram matrix modules for visual emotion classification.
- Propose Gram matrix modules that perform layer-specific dimensionality reduction, to capture compact style related feature representations at multiple depths of a CNN.
- Evaluate the influence of shallow feature enhancement on the accuracy and interpretability of the model.
- Visualize and analyze internal representations to gain qualitative insight into model behavior.

Figure 4.1 illustrates the proposed CNN-based model for VER. The input is an image that conveys an emotion. The EfficientNetV2 convolutional neural network is used for feature extraction. The EfficientNetV2 intermediate feature maps are passed into the corresponding Gram matrix modules. Gram matrix modules capture color, texture, and style features, which are low-level features important for VER.

The extraction of intermediate feature maps from the backbone can be defined as retrieving outputs from selected layers. These feature maps represent learned visual patterns at varying abstraction levels, which is why the backbone is commonly referred to as a feature extractor. Formally, denote the set of extracted feature maps as follows:

Let the convolutional backbone be B with layers $\{l_1, \ldots, l_L\}$. Define an index set $S \subseteq \{1, \ldots, L\}$ of layers from which to extract feature maps, and let $v = |S|$ denote the number of selected layers. For each selected layer $l_i \in S$, let

$$F_{l_i} \in \mathbb{R}^{C_i \times H_i \times W_i} \tag{4.1}$$

denote the corresponding extracted feature map, where C_i, H_i, and W_i denote the number of channels, spatial height, and spatial width of the feature map extracted from layer l_i, respectively.

Each extracted feature map F_{l_i} is then passed to a trainable Gram matrix module $g_i(\cdot)$ (the module is further detailed in Figure 4.3), which computes a feature vector:

$$z_i = g_i(F_{l_i}), \qquad z_i \in \mathbb{R}^{d_i}. \tag{4.2}$$

Here, d_i is the dimensionality of the embedding, defined as

$$d_i = \lfloor C_i/2 \rfloor, \tag{4.3}$$

and z_i is the output feature vector of the Gram matrix module corresponding to layer l_i.

The number of Gram matrix modules is v. The outputs of all Gram matrix modules are then concatenated into a single feature vector:

$$z = [z_1; z_2; \ldots; z_v] \in \mathbb{R}^D, \qquad D = \sum_{i=1}^{v} d_i. \tag{4.4}$$

Let $h \in \mathbb{R}^{C_*}$ denote the global feature vector produced by the backbone after global average pooling, where C_* is the dimensionality of the backbone output.

To make the backbone output compatible with the Gram matrix modules concatenated output, we define a learnable projection (in our case implemented as a fully connected layer)

$$P : \mathbb{R}^{C_*} \to \mathbb{R}^D, \tag{4.5}$$

and define the projected backbone vector as

$$p = P(h), \qquad p \in \mathbb{R}^D. \tag{4.6}$$

The two representations are fused by element-wise addition:

$$u = p + z, \qquad u \in \mathbb{R}^D, \tag{4.7}$$

where u is the fused penultimate feature vector.

Finally, the classifier maps the fused representation u to the output prediction vector

$$\hat{y} = \text{Classifier}(u), \qquad \hat{y} \in \mathbb{R}^N, \tag{4.8}$$

where N is the number of visual emotion classes and $\hat{y}$ denotes the predicted logit vector. To obtain class probabilities from the output logit vector $\hat{y}$, the softmax function is applied to these logits.

The selected backbone feature maps have channel dimensions $C_1 = 48$, $C_2 = 64$, and $C_3 = 160$. Using the described reduction rule $d_i = \lfloor C_i/2 \rfloor$, we obtain $(d_1, d_2, d_3) = (24, 32, 80)$ and the concatenated vector sized $v = 3$ of the outputs of all Gram matrix modules has dimension of $D = 136$.

In Figure 4.2, h denotes the backbone output after global average pooling, p is the projected backbone feature vector, $z_1, \ldots, z_v$ are the outputs of the corresponding Gram matrix modules, z is the concatenated vector of the outputs of all Gram matrix modules, u is the fused penultimate layer feature vector, and $\hat{y}$ is the final classifier output.

The EfficientNetV2S convolutional neural network is utilized as a backbone for feature extraction due to its low number of network parameters and demonstrated good results in the ImageNet [66] benchmark. After global average pooling, the EfficientNetV2 backbone produces a 1792-dimensional feature vector. The EfficientNetV2S model comprises 91 convolutional layers, organized into 6 previously described stages. The dimensionalities illustrated in Figure 4.2 correspond to the specific EfficientNetV2S instance used in the subsequent experiments. In the general case, the backbone output dimension C_* and the layer specific

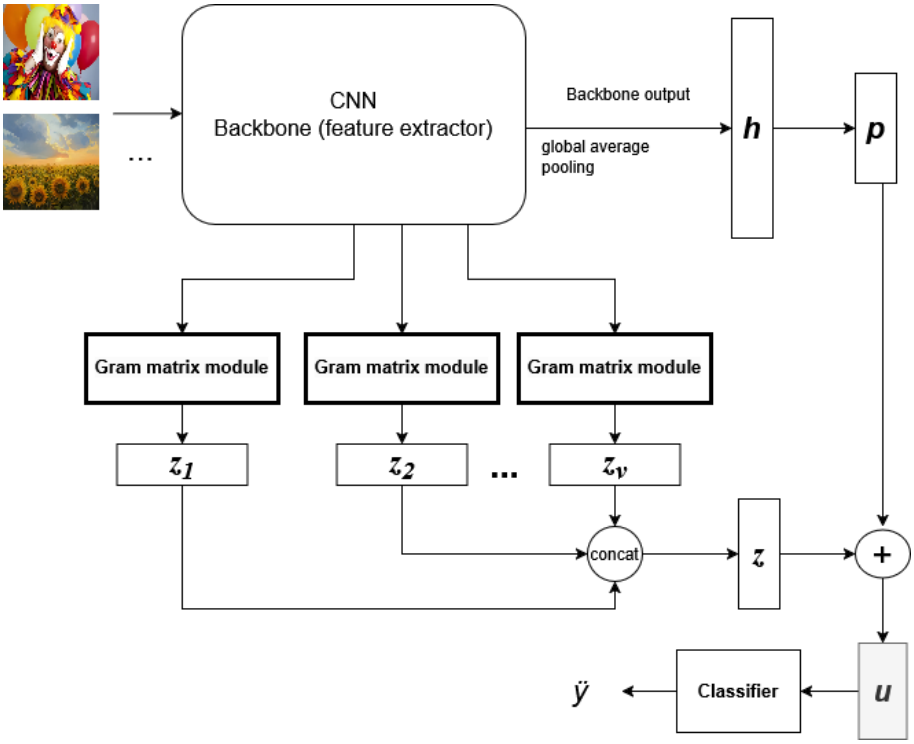


Figure 4.2: Overall scheme of the proposed model.

dimensions d_i depend on the selected architecture and chosen intermediate layers. The proposed CNN-based model design does not constrain these values.

Figure 4.3 shows the Gram matrix module structure. In contrast to our paper [51], we reduced the final output dimensionality of the module, aiming to decrease the number of network parameters. Each module receives input, whose shape is $\mathbb{R}^{C\times H\times W}$, corresponding to the extracted layer's feature map, which consists of C feature sub-maps. Feature sub-maps are defined by their spatial dimensions, $H \times W$, in height and width. It should be noted that the output of the Gram matrix is in a quadratic form and is expressed as $C \times C$ squared matrix. The Gram matrix is flattened into a one-dimensional vector consisting of $C \times C$ units, which is further compressed by a dense layer to $C/2$ units. These units are then applied to an activation function and batch normalization. Here, $C/2$ means dividing the number of obtained original channels by a factor of 2. SiLU activation function is called the sigmoid linear

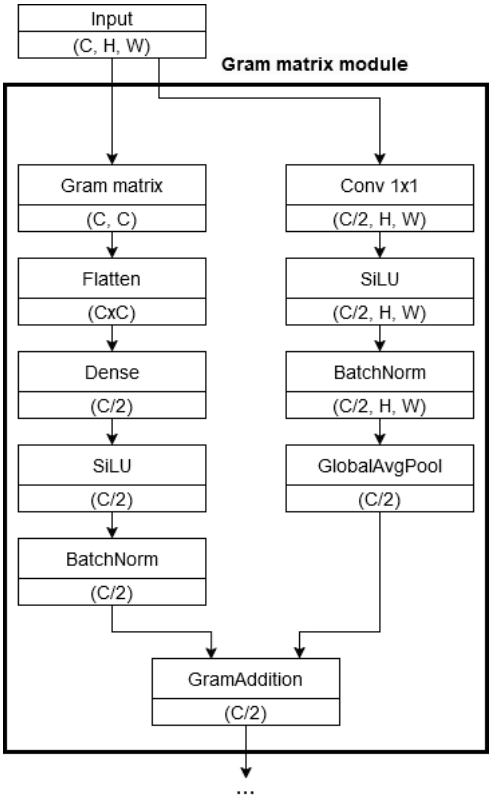


Figure 4.3: Scheme of the auxiliary Gram matrix module.

unit [21], or more commonly known as a swish activation function. The SiLU activation function was used due to its smooth behavior and its ability to produce negative outputs, which helps maintain gradient flow during backpropagation. Accordingly, the other side of the Gram matrix module consists of a 1x1 convolution operation. From this convolution, we get $C/2$ feature sub-maps. In a theoretical case where the number of sub-maps C is odd, we use a floor operation $C/2$ to ensure that reduced dimension is an integer. 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 4.3. We also considered the concatenation, multiplication, and average fusion strategies, but these options increased the number of trainable

parameters and yielded no gains.

The Gram matrix $G \in \mathbb{R}^{C \times C}$ can be written as follows:

$$G = FF^T, \quad F \in \mathbb{R}^{C \times HW}, \quad F^T \in \mathbb{R}^{HW \times C} \tag{4.9}$$

where C denotes the number of channels and H and W represent the spatial height and width of the feature map. Although computed from a single feature map layer, the Gram matrix aggregates information across all spatial locations, providing a global (layer respective) summary of channel relationships.

The Gram matrix values from the deeper feature maps can get a higher value and cause numerical instability. The common solution for this case is to normalize Gram matrix values:

$$G = \frac{FF^T}{H \cdot W}, \quad F \in \mathbb{R}^{C \times HW}, \quad F^T \in \mathbb{R}^{HW \times C} \tag{4.10}$$

In Eq. (4.10) the Gram matrix is divided by $H \cdot W$ height and width spatial dimensions. This normalization effectively reduces the range of values and minimizes the risk of numerical overflow.

Figure 4.4: Illustrative visualizations of Gram matrix–based representations. Top row: original images; bottom row: Gram matrix-based representations displaying texture and color patterns. Original images are from EmoSet-118K dataset.

Figure 4.4 provides an illustrative example of how Gram matrix-based representations encode visual information. The top row shows

the original input images, while the bottom row visualizes the corresponding representations derived from Gram matrices, following the interpretation commonly used in earlier studies on Gram matrix based feature analysis [23].

Although these visualizations are not generated by the proposed model, they offer potential insight into the type of information captured by Gram matrices. By aggregating feature responses across spatial locations, Gram matrices capture texture, color distribution, and stylistic patterns while restraining spatial structure. As a result, the bottom row representations highlight low-level visual features and coarse structural outlines rather than exact object layout.

This property is relevant for visual emotion recognition, where emotional perception is influenced by global appearance, color composition, and texture rather than the exact layout of the object. These examples help to understand the integration of Gram matrix based representations into the proposed architecture.

4.2 METRICS AND EVALUATION CRITERIA

To assess the efficiency of the models and their ability to identify visual emotions, the following metrics are used throughout this dissertation: accuracy, F1 score, precision, and recall. These metrics are calculated as follows:

$$\begin{aligned} Accuracy &= \frac{TP+TN}{TP+TN+FP+FN} \\ Precision &= \frac{TP}{TP+FP} \\ Recall &= \frac{TP}{TP+FN} \\ F1 &= \frac{2 \times Precision \times Recall}{Precision + Recall} \end{aligned} \tag{4.11}$$

Here, TP (true positives) are instances correctly identified as belonging to the positive class, and TN (true negatives) are instances correctly classified as belonging to the negative class. FP (false positives) are negative instances incorrectly labeled as positive, while FN (false negatives) are positive instances incorrectly predicted as negative. Accuracy

measures the overall proportion of correct predictions. The F1 score, which balances precision and recall, is particularly useful when the class distribution is imbalanced. These metrics are sufficient to provide a reliable assessment of model performance. Since emotion recognition is a multiclass task, precision, recall, and F1 are computed using macro-averaging unless stated otherwise. Macro-averaging is performed by computing precision, recall, and F1 separately for each class, treating corresponding class as the positive class in turn, and then taking the unweighted mean of the resulting per-class metric values.

4.2.1 Consistency Measure

In addition to standard classification metrics, we also suggest a measure, the top-2 cross-sentiment measure, which measures the internal consistency of the model's top probabilistic predictions. Unlike accuracy-based metrics, which compare predictions against ground-truth labels, this metric evaluates whether the model assigns high confidence to visual emotion classes belonging to opposite sentiment groups (positive versus negative). A low cross-sentiment rate indicates coherent sentiment behaviour, where a high rate suggests uncertainty or inconsistency in the learned representation.

To compute this measure, we consider the model's top-2 predicted classes for each sample, i.e., the two classes with the highest predicted probabilities. This essentially measures the predicted classes in the top-2 per sample (i.e., the two classes with the highest probability according to the model). The EmoSet-118K and FI-8 datasets are grounded in Mikel's wheel of emotions [49]. This means that we can determine whether the two top predictions belong to the same sentiment group or different ones. We can formally describe this measure in the following way. For each input sample (image) $\mathbf{x}_i$, let (c_1, p_1) and (c_2, p_2) denote the indices of the predicted visual emotion classes and their associated probabilities for the top-1 and top-2 predictions. Define a mapping

$$f : \{0, 1, \ldots, 7\} \rightarrow \{\text{positive}, \text{negative}\},$$

where classes $\{0, 1, 2, 3\}$ are grouped as "positive" and $\{4, 5, 6, 7\}$ are

grouped to "negative", respectively.

A sample is said to be cross-sentiment in its top-2 predictions if $f(c_1) \neq f(c_2)$. Let $\theta \in [0,1]$ denote a confidence threshold. We consider only those samples for which both top-1 and top-2 confidences satisfy

$$p_1 \geq \theta \quad \text{and} \quad p_2 \geq \theta. \tag{4.12}$$

If N_θ denotes the number of such eligible samples, then the top-2 cross-sentiment measure (rate) is defined as

$$\text{CrossRate}(\theta) \;=\; \frac{1}{N_\theta} \sum_{i=1}^{N_\theta} \mathbf{1}[\, f(c_{1i}) \neq f(c_{2i}) \,], \tag{4.13}$$

where $\mathbf{1}[\cdot]$ is the indicator function. There are a few interpretations of this measure:

- If the cross-sentiment rate is high, the model appears to be unsure about the polarity of emotions. This can indicate the model's unreliability,

- If it is low, then the model tends to reliably separate sentiments - a sign of internal consistency.

One might question the need to introduce a confidence threshold. The aim of the threshold is not to capture low-confidence uncertainty, but to identify cases where the model assigns high confidence to two competing classes. These cases represent high-confidence ambiguity, where the model simultaneously assigns a high probability to two classes that belong to opposite sentiment groups. By applying the threshold to the top-2 probabilities, we limit the analysis to instances where ambiguous predictions are meaningful. This excludes low-confidence predictions that do not meaningfully reflect structured ambiguity.

4.2.2 Representation and Feature Space Evaluation Metrics

One of the primary goals of our study is to investigate the trained model, whose structure is described in Figure 4.2. By encouraging intra-class compactness and inter-class separation, the contrastive-center loss [59] is designed to enhance cluster cohesion and separation. We evaluate whether this proposed improvement aligns with empirical clustering quality and classification accuracy. Using cluster metrics, we aim to assess whether the model's feature space learns a meaningful structure – specifically, whether the contrastive-center loss successfully encourages such organization. In this context, a meaningful structure refers to feature embeddings that form compact intra-class clusters and are distinctly separated across different emotion categories. To achieve this, we analyze emotion representations in the high-dimensional feature space of the trained network and suggest the following possibilities:

- Visualizing feature embeddings in 2D space using dimensionality reduction methods, e.g., UMAP [47].
- Cluster analysis, e.g., using k-means, and evaluation of clustering quality by using metrics such as adjusted Rand index (ARI) [27], normalized mutual information (NMI) [27], and ambiguous sample ratio (ASR) [14].

Visualization and cluster analysis are applied to the set of feature vectors extracted from the penultimate layer of the trained model when evaluated on the test dataset. In both cases, the high-dimensional feature outputs from the same penultimate layer serve as the input for analysis. This enables evaluation of the CNN's performance on the test data, both visually and in terms of similarities and dissimilarities between visual emotion. Different cluster quality metrics are used to assess the resulting clusters from multiple perspectives.

ARI [79] 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.

Let $U = \{U_1, \ldots, U_r\}$ denote the partition obtained by the ground-truth labels and $V = \{V_1, \ldots, V_s\}$ the partition obtained by k-means clustering the learned feature vectors, where both partitions group the same set of N samples. A partition here means a grouping of samples into non-overlapping subsets.

The adjusted Rand index (ARI) measures the agreement between the clustering assignments and the ground-truth labels while correcting for chance. It is defined as

$$\mathrm{ARI}(U,V) = \frac{\mathrm{RI}(U,V) - \mathbb{E}[\mathrm{RI}(U,V)]}{\max(\mathrm{RI}(U,V)) - \mathbb{E}[\mathrm{RI}(U,V)]}, \tag{4.14}$$

where U denotes the partition computed by the ground-truth labels, V denotes the partition computed by clustering the learned feature vectors (in Figure 4.2 this is denoted by u), and $\mathrm{RI}(U,V)$ is the Rand index measuring agreement between the two partitions. The term $\mathbb{E}[\mathrm{RI}(U,V)]$ means the amount of clustering agreement expected only due to only random assignment. The expected Rand index means the amount of agreement that appears by random chance, even if no meaningful clustering occurs. So this expected Rand index agreement is subtracted to account for random clustering chance. Higher ARI values indicate better agreement between the clustering structure and the ground-truth labels.

The normalized mutual information (NMI) measures how much information is shared between the partitions U and V. It is defined as

$$\mathrm{NMI}(U,V) = \frac{2I(U;V)}{H(U)+H(V)}, \tag{4.15}$$

where $I(U;V)$ denotes the mutual information between the two partitions, and $H(U)$ and $H(V)$ denote the corresponding entropies. In plain terms, NMI measures how informative the obtained clustering is about the true class structure. Higher NMI values indicate better alignment between the learned feature-space clustering and the ground-truth labels.

The ambiguous sample ratio (ASR) can be defined as follows. Let

$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ be a set of N testing samples, and let there be Q clusters with known centers $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_Q$. Assume $d(\cdot, \cdot)$ is a chosen distance metric, and $\delta > 0$ is a predefined distance threshold.

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

$$\text{Ambiguity}(\mathbf{x}_k) = \begin{cases} 1 & \text{if } \sum_{j=1}^{Q} \mathbf{1}\big[d(\mathbf{x}_k, \mathbf{q}_j) < \delta\big] > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{4.16}$$

Here, $\mathbf{1}[\cdot]$ denotes the indicator function, which equals 1 if the condition inside is satisfied and 0 otherwise.

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

$$\text{ASR} = \frac{1}{N} \sum_{k=1}^{N} \text{Ambiguity}(\mathbf{x}_k). \tag{4.17}$$

A sample $\mathbf{x}_k$ is considered ambiguous if it is close – within distance threshold δ – to more than one cluster center. The ASR is merely the ratio of all samples that meet the chosen ambiguity criterion and the total number of samples. 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 that is the greatest distance among the 10 percent smallest distances. This heuristic focuses the criterion on samples located near potential cluster boundaries. While approximate, this threshold selection strategy provides a practical criterion for detecting samples that lie close to multiple cluster centers. This approach closely resembles methods in fuzzy clustering or probabilistic models (e.g., Gaussian mixture models), which quantify ambiguity using continuous membership values [80], [77], [14].

In the experiments below, we utilized k-means clustering and Set the number of clusters equal to the number of prediction classes from the corresponding dataset.

The described measures enable us to further investigate the behavior of the trained model. By quantifying the model's output, we can gain a

better understanding of visual emotion recognition.

Although ARI, NMI, and ASR measures, as well as embedding visualizations, do not directly evaluate the classification performance of the model, they provide complementary insights into the structure of the learned feature space. Models with higher classification accuracy are expected to learn more coherent, compact, and better-separated emotion representations. These representation-level metrics help assess whether the internal embeddings align with the model's behavior and the expected improvements introduced by contrastive-center loss. This assumption aligns with findings in representation learning, where enhanced emotion recognition performance is often associated with more structured or semantically emotion-aware embedding spaces [52], [87], [88].

4.3 CONTRASTIVE-CENTER LOSS

4.3.1 Motivation

In current developments of visual emotion recognition models, the frameworks 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 methods 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 CNN, a model that uses convolutional operations to extract hierarchical spatial features to derive robust feature representations for emotion recognition.

Our contribution involves integrating contrastive center loss optimization into the training of a deep neural network. We aim to achieve

several improvements in 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 visual emotion classes further apart by leveraging class centers in the feature embeddings.

We investigate the emotion representations of the trained network in the high-dimensional feature output space and gain additional insights into the trained model by employing dimensionality reduction methods. In particular, we use the uniform manifold approximation and projection UMAP [47] dimension reduction technique to visualize the high-dimensional feature outputs. These visualizations enable us to understand positional embeddings, including 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.

We have a penultimate layer output in our model, given in Figure 4.2. The penultimate layer is the network's second-to-last layer, highlighted in greyscale in Figure 4.2. The output of this layer is a 136-feature vector. This feature vector is used to compute the contrastive-center loss. In our case, 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 4.2) is achieved through an auxiliary module.

The contrastive-center loss [59] is formulated as an objective function that is minimized during network training to enhance feature discriminability and is defined as follows:

$$L_{\text{contr}} = L_{\text{intra}} + L_{\text{inter}}. \tag{4.18}$$

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.

Intra-class compactness component is as follows:

$$L_{\text{intra}} = \frac{1}{N} \sum_{k=1}^{N} \|\mathbf{x}_k - \mathbf{c}_{y_k}\|^2, \tag{4.19}$$

where $\mathbf{x}_k$ is the feature vector of the k-th sample (image), and $y_k \in \{1, \ldots, C\}$ denotes the ground-truth class label of the k-th sample. The vector $\mathbf{c}_{y_k}$ is the center corresponding to class y_k. The total number of classes is C, and the class center are denoted by $\mathbf{c}_1, ..., \mathbf{c}_C$. Each center of the class is a learnable vector, initialized randomly and updated via gradient descent during training. The centers are model parameters (such as weights in a neural network) that are optimized alongside the rest of the network during training.

This loss component aims to minimize 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:

$$L_{\text{inter}} = \frac{1}{C(C-1)} \sum_{i=1}^{C} \sum_{j \neq i} \max(0, m - \|\mathbf{c}_i - \mathbf{c}_j\|)^2, \tag{4.20}$$

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 Euclidean distance between the centers of classes i and j. $\max(0, \cdot)$ indicates that no penalty is being 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 that are 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 as a hyperparameter requiring additional tuning.

We suggest extending the loss function

$$L_{\text{contr}} = L_{\text{intra}} + \lambda \cdot L_{\text{inter}}, \tag{4.21}$$

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

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

The selection of λ (which controls inter-class separability in L_{inter}) was evaluated empirically using the FI-8 dataset. We observed weak performance for $\lambda < 1$, where values $\lambda > 5$ led to overfitting and unstable training results. Based on these observations, $\lambda = 1$ was chosen as a stable compromise that balances intra-class compactness and inter-class separation. Dataset specific tuning of λ is possible, but this increases the risk of overfitting the method to individual datasets and reduces comparability across experimental settings. Therefore, $\lambda = 1$ was fixed for all datasets to ensure methodological consistency and to evaluate the general effectiveness of the proposed contrastive-center loss integration.

4.3.2 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 study [102] 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; its definition has been previously established in Eq. (3.1).

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

$$L_{\text{total}} = L_{\text{entr}} + \beta \cdot L_{\text{contr}}, \tag{4.22}$$

where L_{entr} is the computed sparse categorical cross-entropy loss from the main layer, and L_{contr} is the contrastive-center loss from the penultimate layer. The coefficient β is a weighting hyperparameter that controls the relative importance of the contrastive-center loss in the total loss function.

4.4 EXPERIMENTAL SETUP

This section describes the experimental setup used to implement and evaluate the proposed model for VER.

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) using a momentum of 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 [43]: it decays from 0.02 down to 1×10^{-4} and restarts back every five epochs. The parameter β allows us to control 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 [13] with

distortion strength set to three, and number of transformations set to two.

The integration of the contrastive-center loss (refer to subsection 4.3.2) 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 [25]. To mitigate this, we employed gradient clipping (normalization). Specifically, given the original gradient vector g, the clipped gradient g_{clipped} is computed as

$$g_{\text{clipped}} = g \cdot \min\left(1, \frac{h}{\|g\|_2}\right), \tag{4.23}$$

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 $h = 5$. Initial results showed that smaller thresholds ($h < 3$) led to aggressive gradient scaling and slower convergence, where larger values did not successfully reduce gradient spikes introduced by the contrastive-center loss term. The selected value provided stable training results.

The described model was trained using the PyTorch deep learning library. Initially, the first model versions were trained on the TensorFlow deep learning library. It is possible to achieve consistent model results with fixed weight initialization using PyTorch, whereas this is not the case for TensorFlow [2]. This allows for achieving matching model performance between different training iterations without underlying uncertainty. In practice, we still expect the following results to have uncertainty.

The PyTorch library has an automatic mixed precision (AMP) feature. Usually, the models are trained using 32-bit floating point precision (FP32). Newer generation hardware has the capability to utilize computations using 16-bit floating point precision (FP16). Therefore, the PyTorch supports this capability to automatically choose the proper precision for the corresponding network parts. The following network parts are kept in FP32 [48] [3]:

- Master copy of model weights, the optimizer maintains the copy

in order to maintain precision drift.

- Loss value and its scaling requires good dynamic range to avoid underflow.
- BatchNorm – running mean/variance are kept in FP32.

The following model parts are kept in FP16:

- Forward pass activations – convolutions, matrix multiplications.
- Backward pass activations – gradients with respect to activations.

The most significant drawback of AMP is the risk of numerical instability. FP16 has a lower numerical range and its easier to get onto overflow and underflow, which in turn produces not a number values (NaNs). Even newer hardware supports 16-bit brain floating point (BF16). The main distinction from FP16 is that BF16 keeps full FP32 exponent range with a lower mantissa precision. Many training pipelines prefer AMP with BF16 setup. The training pipeline is summarized in the following order;

- Convert the input images to the colored 224x224 pixel space.
- Apply the RandAugment augmentation technique with values set to $M = 3$ and $N = 2$. The given values are for the magnitude of augmentation transformations and N refers to the number of applied augmentations to the input image.
- Evaluate and determine feasible hyperparameter values.

4.5 CONCLUSIONS OF THE CHAPTER

This chapter presents a CNN-based model and training strategy for visual emotion recognition in general-purpose images. The proposed model extends a CNN backbone with Gram matrix modules that capture complementary low-level visual features from intermediate feature

maps. These modules are designed to capture stylistic attributes such as color, texture, and repeating visual patterns, and to integrate them with the semantic features learned by the backbone. The developed model aims to retain stylistic information that standard CNN pipelines often overlook, enabling more accurate emotion classification.

In addition, this chapter introduces contrastive-center loss as an additional loss function component for improving emotion class discriminability. By encouraging intra-class compactness and enforcing inter-class separation through learnable class centers and a margin constraint, the proposed loss addresses a gap of standard cross-entropy-based training for similar visual emotion classes. The proposed training strategy also defines practical training components needed for stable optimization, including tuning hyperparameters, gradient clipping, and Gram matrix normalization to avoid numerical instability.

Finally, the chapter presents additional evaluation methods for analyzing learned emotion representations. Dimensionality reduction visualizations, cluster-quality measures (ARI, NMI, ASR), and the top-2 cross-sentiment measure provide additional means of assessing the internal structure of the learned feature space. These evaluations enable both qualitative and quantitative analysis of whether the proposed model and training strategy lead to more coherent, compact, and better-separated emotion representations, forming the basis for the experimental validation presented in the next chapter.

5 EXPERIMENTS AND RESULTS

This chapter presents an evaluation of the proposed CNN-based model for visual emotion recognition. First, we quantify the contribution of the auxiliary Gram matrix modules by comparing the backbone baseline against configurations with different numbers of parallel Gram matrix modules, using standard classification metrics and utilizing various visual emotion datasets. Then, we evaluate the effect of integrating contrastive-center loss, examining both predictive performance and representation-level structure through clustering metrics (ARI, NMI, ASR) and UMAP visualizations. Finally, we demonstrate practical applicability through case studies on artwork and WikiArt images, including a consistency analysis based on top-2 cross-sentiment measure.

The experiments and results have been partially published in peer-reviewed papers [A1, A2, B1].

5.1 EVALUATING THE GAIN OF GRAM MATRIX MODULES

The aim of the experimental study is to compare the proposed new model with the backbone using the aforementioned metrics and a dataset of annotated images that convey emotions. It is also necessary to determine the appropriate number of Gram modules to be connected in parallel.

In Table 5.1, the averaged accuracy with standard deviations SD is shown. Our proposed network outperforms the baseline network with around 1.2% higher accuracy. Using two Gram matrix feature extraction modules ($v = 2$) yields a slightly better result compared to the case of three modules ($v = 3$). Using four Gram matrix modules produces encouraging results. We also evaluated the model of Zhang et al. [96] using the training setup reported in [96] (60 epochs). Our baseline and Gram matrix module variants were trained for 20 epochs, which was enough for convergence for our training setup. Under these

Table 5.1: Accuracy on the WEBEmo sadness test set results averaged over 3 runs and compared to the baseline. Baseline corresponds to the backbone EfficientNetV2S network.

Network	Accuracy (%)	SD
Baseline	81.308	±0.548
Zhang et al. [96]	81.313	±0.186
2 Gram modules	82.313	±0.239
3 Gram modules	82.171	±0.128
4 Gram modules	82.520	±0.224

respective setups, EfficientNetV2S achieves higher accuracy than Zhang et al [96], where the backbone of their model is ResNet50. Let us note that the standard deviation is smallest when $v = 3$. Here, we can assume that a greater number of Gram modules produces some stability of the results and of the network in general. However, case $v = 4$ gives a lower standard deviation than the baseline model, and accuracy is better.

Table 5.2: Precision, recall and $F1$-score results averaged over three runs with standard deviations. Baseline is the backbone EfficientNetV2S network.

Model	**Others**			**Sadness**		
	Precision	Recall	$F1$-score	Precision	Recall	$F1$-score
Baseline	0.8219 ± 0.0069	0.8359 ± 0.0053	0.8288 ± 0.0047	0.8024 ± 0.0056	0.7862 ± 0.0099	0.7942 ± 0.0066
1 module	0.8336 ± 0.0018	**0.8476 ± 0.0037**	**0.8405 ± 0.0024**	**0.8165 ± 0.0038**	0.8004 ± 0.0023	**0.8084 ± 0.0025**
2 modules	0.8313 ± 0.0036	0.8446 ± 0.0036	0.8379 ± 0.0021	0.8132 ± 0.0032	0.7978 ± 0.0056	0.8054 ± 0.0030
3 modules	0.8312 ± 0.0021	0.8415 ± 0.00053	0.8363 ± 0.00094	0.8102 ± 0.00046	0.7984 ± 0.0032	0.8042 ± 0.0017
4 modules	**0.8353 ± 0.0059**	0.8434 ± 0.0057	0.8393 ± 0.0017	0.8131 ± 0.0041	**0.8037 ± 0.0095**	**0.8084 ± 0.0036**

In Table 5.2, the classification results averaged over three runs are shown. The model was trained on the WEBEmo sadness training set and evaluated on the testing set. Compared to the baseline, all Gram matrix module setups improve the others class precision, recall, and $F1$-score. The highest others $F1$-score is achieved with $v = 1$ module (0.8405 ± 0.0024), while the highest others precision is achieved with $v = 4$ modules (0.8353 ± 0.0059). For the sadness class, all Gram matrix module setups improve precision and $F1$-score over the baseline, and the highest sadness recall is achieved with $v = 4$ modules (0.8037 ± 0.0095).

Overall, the results indicate consistent gains over the baseline across both classes, with the best-performing configuration depending on the chosen metric. The gains are consistent, but the suitable v differs by metric: $v = 1$ maximizes the others $F1$-score, whereas $v = 4$ produces the highest sadness recall.

5.1.1 Applying the Trained Networks on Other Datasets

In this section, we present the results of using our suggested trained networks on the WEBEmo data on other emotion image datasets. Additionally, the trained baseline network's performance is presented for comparison as well. Our proposed network consists of $v = 4$ Gram matrix modules. The networks were trained using WEBEmo sadness dataset as described in sectiom 4.4. Then, we use UnbiasedEmo [61] and Emotion-6 [61] subsets for analysis using the trained networks. The goal of the experiment is to estimate the generalization capability of trained networks on other unseen datasets.

Table 5.3: Testing report of the trained network of 3 averaged runs with $v = 4$ Gram matrix modules using the UnbiasedEmo dataset.

Class	Precision	Recall	$F1$-score
Others	0.8772 ± 0.0008	0.7902 ± 0.0175	0.8313 ± 0.0094
Sadness	0.7548 ± 0.0144	0.8536 ± 0.0043	0.8011 ± 0.0063

Table 5.4: Testing report of the trained network of 3 averaged runs with $v = 4$ Gram matrix modules using Emotion-6 dataset.

Class	Precision	Recall	$F1$-score
Others	0.8407 ± 0.0036	0.6670 ± 0.0192	0.7437 ± 0.0108
Sadness	0.5572 ± 0.0102	0.7679 ± 0.0125	0.6455 ± 0.00445

Tables 5.3 and 5.4 display evaluations on UnbiasedEmo Emotion-6 ($v = 4$) using model trained on the WEBEmo sadness dataset. It can be noted that the proposed model on the particular cases has the highest $F1$-score for the sadness emotion class. On the UnbiasedEmo testing set, the proposed network performs comparatively well across both classes. The network demonstrates reliable results in terms of sadness

class precision and $F1$-score values. However, on the Emotion-6 testing set, the proposed network performs worse in distinguishing sad image emotions. The reason might be that there is a slight class imbalance, where the majority group is highly favored. Interestingly, although the networks were trained on the WEBEmo dataset, their evaluation on the UnbiasedEmo and Emotion-6 datasets (Tables 5.3 and 5.4) shows even higher precision for the others class compared to the WEBEmo sadness testing subset (Table 5.2).

5.2 CONTRASTIVE-CENTER LOSS INTEGRATION RESULTS

This section presents an experimental evaluation of the integration of the proposed contrastive-center loss. The analysis focuses on evaluating the weighting coefficient β influences classification performance and the structure of learned feature representations. Quantitative results are reported using standard classification and through cluster quality metrics across multiple emotion datasets, followed by qualitative visualization of the feature space to provide additional insight into the effects of the contrastive-center loss on the emotion separability.

5.2.1 Comparison of Metrics

The aim of the experimental study is 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. Here, the baseline and the proposed model are formulated as described in Figure 4.2 with $v = 3$ Gram matrix modules.

Table 5.5 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 4.3.2). Case $\beta = 0$ corresponds to the baseline model, where the contrastive-center loss is not used and training is performed only using the sparse categorical cross-entropy loss defined in Eq. (3.1). Each row reports the performance of a trained model on a binary emotion classification

Table 5.5: 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
1.2	0.8284	**0.1402**	**0.2198**	**0.2801**

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 5.6: Performance metrics on dependence on β; the results are from evaluating the 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
1.2	0.7138	0.4682	**0.4755**	**0.0728**

Table 5.6 shows the performance of the network using the FI-8 dataset, where training was performed on the training set and evaluated on the testing set. 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, and lowest ASR were achieved with $\beta = 1.2$. However, accuracy improvement is marginal. In our case, we can observe that clustering quality improves when integrating the contrastive-center loss optimization. This suggests that the model gains the capability to recognize visual emotion classes more effectively.

Table 5.7: Performance metrics on dependence on β; the results are from evaluating the 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
1.2	**0.7886**	0.5093	0.5906	**0.0759**

Table 5.7 shows the performance of the network using the EmoSet-118K 8-dataset, where training was performed on the training set and evaluated on the testing set. 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 $\beta = 1.2$. Overall, these results confirm that integrating the contrastive-center loss enhances feature clustering. This indicates that the trained network performs better in tasks with a larger number of visual emotion classes.

Table 5.8 shows the performance of the network using the CAER-S dataset, where training was performed on the training set and evaluated on the testing set. The network was trained .to recognize emotion from seven classes. The highest accuracy was obtained for $\beta = 1.0$. In comparison to the baseline, integrating contrastive-center loss optimization did not yield a substantial improvement in accuracy. The highest ARI and NMI scores were achieved with $\beta = 0.3$. However, differences are

Table 5.8: Performance metrics on dependence on β; the results are from evaluating the CAER-S testing set.

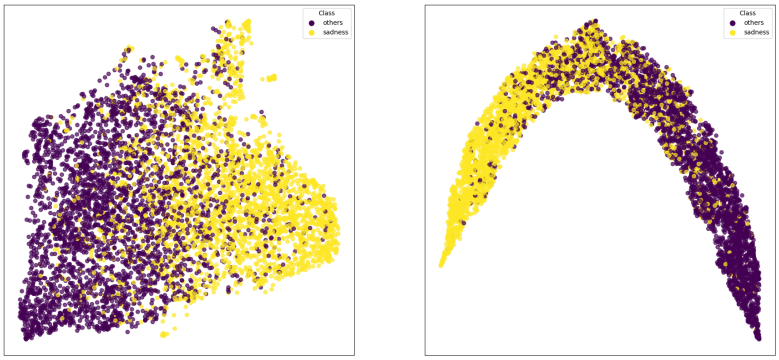
β	Accuracy	ARI	NMI	ASR
0.0	0.9104	0.7139	0.7438	0.7963
0.1	0.9106	0.8100	0.7880	0.0547
0.3	0.9118	**0.8142**	**0.7919**	0.0227
0.4	0.9114	0.7103	0.7425	0.0083
0.5	0.9106	0.7289	0.7671	0.0041
0.8	0.9112	0.7082	0.7533	0.0034
1.0	**0.9120**	0.7295	0.7670	0.0007
1.2	0.9104	0.7264	0.7673	**0**

fairly marginal.

Table 5.9: Performance metrics on dependence on λ; $m = 5$, $\beta = 0.5$; the results are from evaluating the WEBEmo sadness testing set.

λ	Accuracy ↑	ARI ↑	NMI ↑	ASR ↓
0.05	0.8278	0.1317	0.2230	0.3229
0.10	0.8306	0.1594	0.2245	0.0884
0.60	0.8335	0.1758	0.2242	0.0573
1.00	0.8332	0.1708	0.2279	0.0624
2.00	0.8369	0.1680	0.2280	0.0650
5.00	**0.8371**	0.1742	0.2279	0.0470
10.00	0.8335	**0.1872**	**0.2292**	**0.0255**
20.00	0.8350	0.1783	0.2263	0.0452

Table 5.9 presents the performance of the model under a fixed margin $m = 5$ and $\beta = 0.5$, while varying the inter-class weighting hyperparameter λ (see Eq. 4.21). The results are presented evaluating the WEBEmo sadness testing set, while the model was trained on the WEBEmo sadness training set. The results indicate that accuracy changes only slightly once λ becomes sufficiently large, where clustering quality metrics continue to improve with increasing λ. This suggests that larger λ values primarily enhance class separability in the embedding space and reduce prediction ambiguity, rather than resulting in massive gains in classification accuracy.



(a) Contrastive-center loss weighting coefficient $\beta = 0$.

(b) Contrastive-center loss weighting coefficient $\beta = 0.5$

Figure 5.1: Visualization of the trained model results from the WEBEmo sadness testing set.

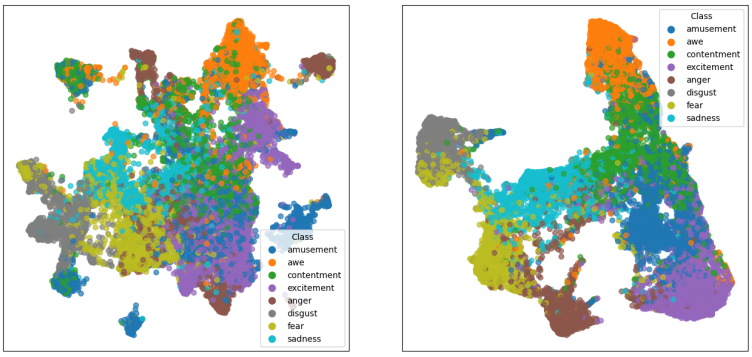
The results demonstrate consistent improvements in clustering metrics when integrating the contrastive-center loss. Improvements in classification accuracy are also notable across datasets. This behavior is expected in visual emotion recognition, which still remains a challenging problem. The results indicate that the primary effect of the contrastive-center loss lies in improving the structure and separability of the learned feature representations rather than significantly improving classification accuracy. To better understand how the feature space is enhanced by the contrastive-center loss, a qualitative analysis based on visualization of the penultimate layer embeddings is presented next. This analysis provides insight into changes in class compactness, separation, and overlap that are not directly visible by performance metrics.

5.2.2 Visual Analysis

Let's 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 is equal to the number of images in the set of test images. Let us use, e.g., the UMAP method [47] 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 [51]. The results are given in Figure 5.1. 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 visualized 6108 vectors, corresponding to the 6108 emotion images in the testing subset. A single point in the figure represents the testing image on a plane, and the coloring indicates the ground-truth labels – the true class, not predicted by the network – of the corresponding image.

We can observe that there is no clear, distinct boundary between the two emotion groups. However, we observe a polarity (contrast) of the distribution of points on the plane: visual emotion classes tend to form apparent groups as seen in color distribution in Figure 5.1. A very interesting discovery is that in case (b) the distribution of points on a plane has a 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.

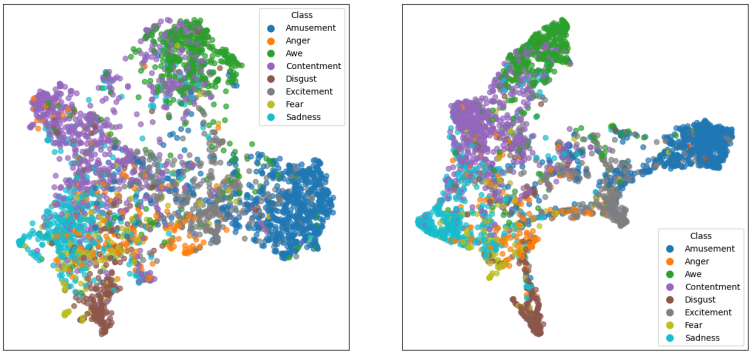


(a) $\beta = 0$ (b) $\beta = 0.5$

Figure 5.2: Visualization of the trained model results from the EmoSet-118K testing set.

In the second experiment, the network is trained on the EmoSet-118K dataset. The results are given in Figure 5.2. Two models were used: (a) baseline, $\beta = 0$, and (b) integrated contrastive-center loss optimization, $\beta = 0.5$. In both cases, we visualized 17,716 vectors, corresponding to the 17,716 emotion images in the testing subset. A single point in the

figure represents the testing image on a plane, and its color corresponds to the ground-truth labels. There are eight classes: anger, disgust, fear, sadness, amusement, awe, contentment, and excitement. We observe clusters of emotions in (a). In addition to the clusters, we see some polarization of positive and negative emotions in (b). Therefore, a subtle separation exists between the negative and positive emotion groupings in the plot. Some overlap of classes is apparent both in (a) and (b), but in (b) it is much less.



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

Figure 5.3: Visualization of the trained model results from the FI-8 testing set.

In the third experiment, the network is trained on the FI-8. The results are given in Figure 5.3. Two models were used: (a) baseline, $\beta = 0$, and (b) integrated contrastive-center loss optimization, $\beta = 0.5$. In both cases, we visualized 3,407 vectors, corresponding to the to the same number of emotion images in the testing subset. A single point in the figure is testing image representation on a plane, and the coloring indicates ground-truth emotions. The same total number of classes is eight: anger, disgust, fear, sadness, amusement, awe, contentment, and excitement. We observe clusters of emotions in (a). In addition to the clusters, we see some polarization of positive and negative emotions in (b). Therefore, a subtle separation exists between the negative and positive emotion groupings in the plot. Overlap of classes is apparent both in (a) and (b). Nevertheless, the integration of contrastive-center loss 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 approximately 160 samples for each of these two emotions. Compared to the other classes, there are between 400 and 750 samples for each emotion group, meaning that the trained network still has difficulty identifying minority-class emotions.

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

5.2.3 Distribution of the Class Centers

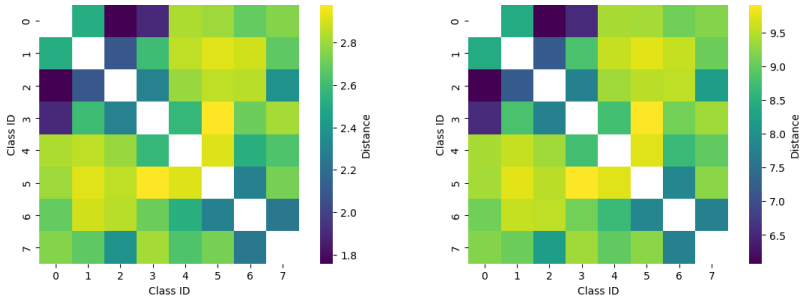
Let us consider a trained network. In Section 4.3, we introduced the term of the class center. The class centers $\mathbf{c}_{y_i}$ are learnable parameters initialized randomly and updated via gradient descent alongside the model parameters during training.

In particular, we are concerned primarily with the inter-class separability loss component L_{inter}. This component has two parts. This component has two parts: first, centers that correspond to the emotion class; and second, a margin m that enforces minimum distances between pairwise centers. It means that every class center must be 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 useful in the additional evaluation of training quality.

In Figures 5.4 and 5.5, 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 $m = 10$.

From both plots in Figure 5.4, we see that the trained model maintains the selected distances sufficiently well with different margins. Several emotion class centers are too close, which leads to gaps in the model's ability to discern those emotions.

This behavior can be further interpreted by inspecting the color scales shown on the right side of the distance matrices. For the case of margin



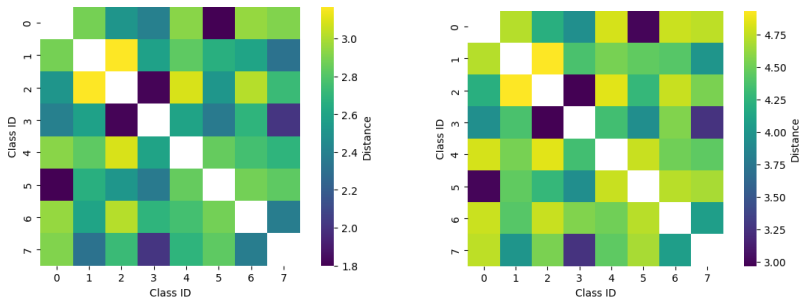
(a) margin $m = 3$ (b) margin $m = 10$

Figure 5.4: Trained model pair-wise center distance matrix. The model was trained on the EmoSet-118K dataset.

$m = 3$, the learned pair-wise center distances mostly lie in the range of approximately 1.8 to 2.9, where for a larger margin $m = 10$, the distances are shifted to a higher range, approximately between 6.5 and 9.5. These ranges directly reflect the distances between the learned contrastive-center class centers after training. Although the margin parameter enforces a minimum distance constraint, which is not fixed, the observed color distributions indicate that the optimization process adapts the inter-class distances in accordance with the selected margin value. The same observations also apply to Figure 5.5.

Figure 5.5 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 5.5 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, 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 good responsiveness of the training process to the selected hyperparameter margin m.



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

Figure 5.5: Model pair-wise center distance matrix. The model was trained on the FI-8 dataset.

Another hyperparameter for weighting inter-class separability loss component L_{inter} is λ (refer Eq. 4.21). The best accuracies were obtained with the following λ values: FI-8 case $\lambda = 100$, EmoSet-118K case $\lambda = 100$ and for WEBEmo case $\lambda = 5$. We observe a notable diversity in optimal λ values across different datasets. However, the influence of λ is not very significant when its values grow: it suffices to use λ more or less than five.

Table 5.10: 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**

Table 5.10 displays the summary of the best-performing configurations obtained for each dataset after tuning the contrastive-center loss hyperparameters, including the weighting coefficient β, the margin m, and λ. For each dataset, the reported results correspond to the most

appropriate hyperparameter setting identified through the experiments shown earlier in this section.

Compared to the baseline models, which were trained without the contrastive-center loss, the proposed CNN-based model with $v = 3$ Gram matrix modules consistently improves classification accuracy and the quality of the learned feature representations, as reflected by notably higher ARI and NMI scores and lower ASR values. These results indicate that the selected hyperparameter configurations enable the model to learn more compact and better-separated emotion clusters in the embedding space. The aggregated results confirm that the integration of contrastive-center loss with properly chosen hyperparameters leads to more coherent emotion representations and improved discrimination in multiple visual emotion datasets.

5.3 PRACTICAL USE CASE STUDY

Evaluating artworks is one of the possible practical use cases for a developed visual emotion prediction model. There is a lack of studies that describe the practical application of emotion recognition models. Therefore, it can be a valuable insight to analyze the practical application of image emotion recognition.

5.3.1 Practical Case Study on the Artwork Images

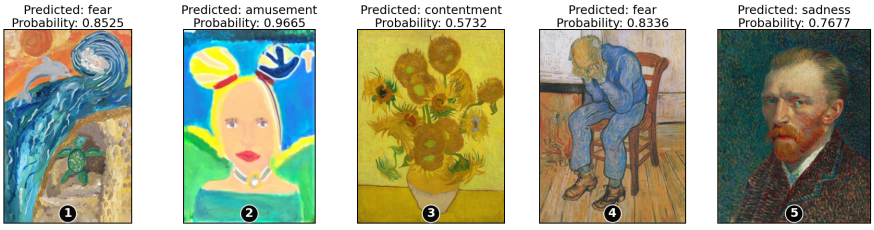


Figure 5.6: Visual emotion recognition on the artwork images.

The aim of this section is to illustrate the recognition of emotions in the images of a general nature. We do not intend to compare our

solution with a baseline; instead, we aim to illustrate the possibilities of the analysis.

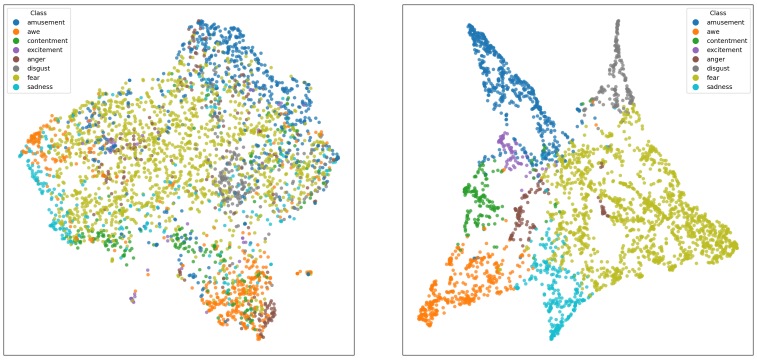
In Figure 5.6, several image artworks conveying emotion are shown. The first two examples correspond to artworks created by a child, while the remaining three are paintings by Vincent van Gogh. In our example, we classify images images using a network trained with contrastive-center integration improvement. Above the pictures, we present the predicted class and the probability of dependence of the picture on the predicted class. It can be noted that the network recognizes some of Van Gogh's artworks as expressing the emotion of sadness. Interestingly, the trained network predicts contentment expression in the third image, which is a painting of a flower bouquet. We might wonder whether the trained network recognizes emotion based on these features: colors, textures, and physical expressions. From the given example, it can be observed that there are common feature details, such as darker colors, texture, and physical shapes, among the images that are recognized as expressing the sadness emotion.

5.3.2 WikiArt Emotions

Another visual emotion dataset is WikiArt emotion [50]. This dataset consists of 4,105 pieces of artwork images that are labelled for emotions evoked towards the observer. We managed to obtain 3,176 images belonging to three sentiment groups: negative, positive, or neutral (mixed).

Figure 5.7 compares the visualization results of the baseline and our model. We can visualize the model's predicted feature vectors, as we did in Section 5.2.2. In this case, 3,176 feature vectors are used in visualization by the UMAP technique. In the case of (b), the distribution of points is better established compared to the baseline case (a). This also indicates that the improvements are consistent across different visual emotion datasets.

As described in the previous chapter, it can be valuable to check the model's internal consistency. We can recall a measure of the model top-2 cross-sentiment measure (described in Eq. 4.13). This experiment aims



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

Figure 5.7: Inference visualization on the unlabeled WikiArt emotions dataset.

to verify whether the improved embedding structure reduces sentiment-level confusion. This sentiment-level analysis does not require ground-truth labels and reveals representational coherence.

Table 5.11: Top-2 cross-sentiment measure results on the WikiArt emotions set (Cross in %).

No threshold	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	Cross
Baseline	3,176	652	664	518	1,342	41.4
Ours	3,176	290	327	891	1,668	**19.4**
With threshold $\theta = 0.3$	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	Cross
Baseline	552	122	121	85	231	43.5
Ours	277	22	33	59	163	**19.9**

In Table 5.11, the top-2 cross-sentiment measure results are shown. Here, a comparison is made between the baseline and the proposed model. In our case, the model conducted classification on the WikiArt emotions dataset. Here, eligible refers to number of samples where both top-1 and top-2 probabilities $\geq \theta$ (if threshold applied). Our proposed model achieves a notably lower cross-sentiment rate compared to the baseline.The cross-sentiment rate reflects the sentiment-level confusion in the model's predictions; lower values indicate better separability of emotion representations. Applying the threshold $\theta = 0.3$ limits the

analysis to the cases where the model assigns confidence to both its top-1 and top-2 predictions. This filters out predictions where the second-best confidence is low, and highlights instances where the model is indeed uncertain between two alternatives.

Table 5.12: Top-2 cross-sentiment measure results on the FI-8 testing set (Cross in %).

No threshold	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	Cross
Baseline	3,407	544	357	1,984	522	26.4
Ours	3,407	418	250	2,097	642	**19.6**
With threshold $\theta = 0.3$	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	Cross
Baseline	489	63	50	298	78	23.1
Ours	331	26	40	208	57	**19.9**

Table 5.13: Top-2 cross-sentiment measure results on the EmoSet-118K testing set (Cross in %).

No threshold	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	Cross
Baseline	17,716	1,710	1,710	8,983	4,726	22.6
Ours	17,716	507	738	10,227	6,244	**7.0**
With threshold $\theta = 0.3$	Eligible	pos-neg	neg-pos	pos-pos	neg-neg	Cross
Baseline	2,330	191	179	1,442	518	15.9
Ours	1,420	73	59	1,022	266	**9.3**

Tables 5.12 and 5.13 present the top-2 cross-sentiment measure results on the FI-8 and EmoSet-118K testing sets. The models were trained on the corresponding visual emotion dataset. In both datasets, the proposed model consistently achieves a notably lower top-2 cross-sentiment compared to the baseline, displaying reduced confusion between positive and negative sentiment groups. This improvement is observed both without threshold and when applying the confidence threshold $\theta = 0.3$, which shows that the proposed model produces more sentiment consistent top-2 predictions for both cases. Overall, the results across both datasets confirm that the modified training strategy integrating contrastive-center loss enhances sentiment-level discriminability and improves internal consistency of the learned representations.

In Figure 5.8, artwork representations are displayed. Artwork im-

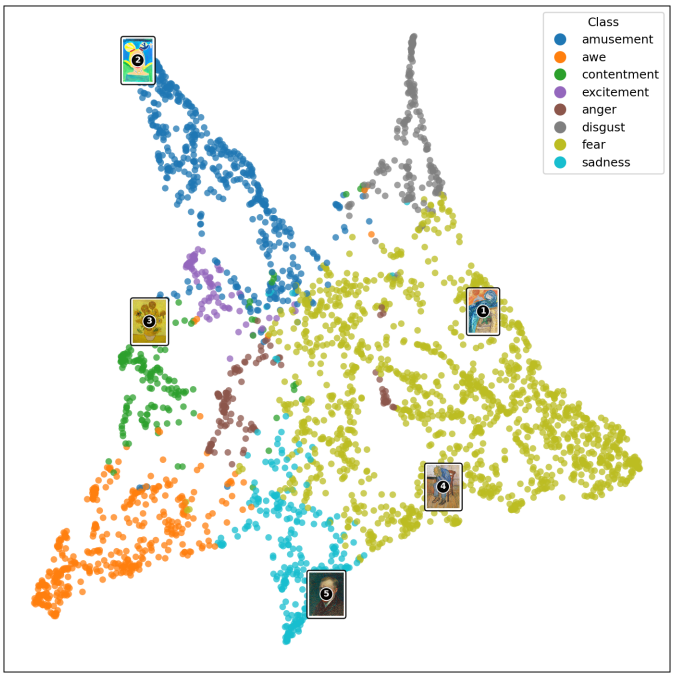


Figure 5.8: Overlay of artworks representation.

ages are the same as in Figure 5.6. In this case, we obtain possible representations of artwork images when the model was trained with contrastive-center integration. The model predicts that children's drawn artwork expresses fear in the first case and amusement in the second case, respectively. Vincent van Gogh's artworks model predicts expressing contentment, fear, and sadness emotions for the corresponding images.

5.4 CONCLUSIONS OF THE CHAPTER

This chapter evaluated the proposed improvements experimentally and demonstrated that adding Gram matrix modules to the EfficientNetV2S backbone improves sadness versus others classification on the WEBEmo sadness dataset. The best configuration utilized four parallel Gram modules, resulting in an increase in accuracy from $81.308\% \pm 0.548\%$ to $82.52\% \pm 0.224\%$ and providing the most stable results. Evaluation

tests on UnbiasedEmo and Emotion-6 indicate that the trained model performs comparatively well on the unseen data.

The chapter also confirmed that integrating contrastive-center loss improves emotion recognition, with accuracy gains across WEBEmo sadness, FI-8, and EmoSet-118K, and with the margin hyperparameter playing an important role (the value $m = 5$ generalized well across the main datasets). Beyond accuracy, representation analysis revealed more structured embeddings: ARI and NMI increased while ASR decreased, and UMAP plots exhibited clearer grouping and polarity separation when $\beta > 0$. Finally, the practical artwork and WikiArt studies illustrated practical use cases, including improved prediction stability evaluated by the top-2 cross-sentiment measure, without requiring ground-truth labels.

6 GENERAL CONCLUSIONS

The dissertation addressed the problem of visual emotion recognition in general-purpose images by developing and evaluating extensions of the convolutional neural network architectures and training methods. The research focused on bridging the affective gap and enhancing class separability, both of which are critical challenges in emotion analysis.

The main conclusions of the dissertation are as follows:

1. Visual data represents a medium for expressing and perceiving emotions. We need automated methods that can effectively interpret large amounts of visual emotion data. This grounds a development of a model architecture and training strategy designed to capture both semantic and stylistic information, improving class separability in the embedding space. By integrating Gram matrix-based feature representations with contrastive-center loss via joint optimization, the proposed CNN-based model aims to achieve more reliable emotion classification in general-purpose images.

2. Adding Gram matrix modules to the EfficientNetV2S backbone improves visual emotion classification accuracy across all tested configurations. The baseline network achieved $81.31\% \pm 0.548\%$, while the proposed model obtained $82.52\% \pm 0.224\%$ with four Gram matrix feature modules, representing an improvement of 1.2% over the baseline on the WEBEmo sadness dataset. The model proposed by Zhang *et al.* [96] achieved $81.313\% \pm 0.186\%$ accuracy, indicating that the proposed approach provides further improvement against the previous Gram matrix-based methods. Furthermore, the four Gram matrix modules option also displayed the most stable performance (lowest standard deviation), indicating improvements in both accuracy and reliability.

3. Integrating contrastive-center loss consistently improved model classification accuracy across all evaluated emotion datasets. On the WEBEmo sadness dataset, accuracy improved from 82.14%

of baseline to 83.74% in the proposed model configuration. On the FI-8 dataset, accuracy improved from 69.68% to 71.88% in the proposed configuration with the best results β around 0.5. On EmoSet-118K, accuracy increased from 77.94% to 80.46% in the proposed model with the best result of $\beta \approx 1$. Overall, the inclusion of joint contrastive-center optimization demonstrates notable improvements across various visual emotion datasets.

4. Additionally, cluster analysis shows that contrastive-center loss improves the structure of the learned feature space. Across image emotion datasets, ARI and NMI increase while ASR decreases, suggesting tighter emotion clusters and fewer samples that belong close to multiple class centers. On EmoSet-118K, ARI increases from 0.332 to 0.5988 and NMI from 0.455 to 0.6241, while ASR drops to 0. Similarly, on the WEBEmo sadness dataset, ASR decreased from 0.8797 to 0.0665. These observations are consistent with dimensionality reductions and embedding visualizations. The proposed model shows clearer groupings and stronger polarity separations when the penalty term $\beta > 0$, supporting the conclusion that contrastive-center optimization yields more coherent and better-separated emotion representations.

5. The proposed top-2 cross-sentiment rate provides a complementary evaluation of prediction consistency by measuring how often the model's two most confident outputs fall into opposite sentiment groups. On the WikiArt emotion dataset, the proposed model produces a notably lower cross-sentiment rate than the baseline both without a confidence threshold (41.4% to 19.4%) and when restricting the analysis to higher-confidence predictions with $\theta = 0.3$ (43.5% to 19.9%). This shows more stable sentiment-level behavior and supports the effectiveness of contrastive–center optimization in reducing ambiguity between opposing emotion groups.

The dissertation thus achieves its goal of proposing and validating effective CNN-based model for visual emotion recognition. Future research nay focus on extending the methodology to multi-modal emotion recognition and analyzing practical applicability in healthcare.

[1] Hojjat Abdollahi, Mohammad H. Mahoor, Rohola Zandie, Jarid Siewierski, and Sara H. Qualls. Artificial emotional intelligence in socially assistive robots for older adults: A pilot study. *IEEE Transactions on Affective Computing*, 14:2020–2032, 7 2023. ISSN 19493045. doi: 10.1109/TAFFC.2022.3143803.

[2] Saeed S. Alahmari, Dmitry B. Goldgof, Peter R. Mouton, and Lawrence O. Hall. Challenges for the repeatability of deep learning models. *IEEE Access*, 8:211860–211868, 2020. ISSN 21693536. doi: 10.1109/ACCESS.2020.3039833.

[3] Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, CK Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Michael Suo, Phil Tillet, Eikan Wang, Xiaodong Wang, William Wen, Shunting Zhang, Xu Zhao, Keren Zhou, Richard Zou, Ajit Mathews, Gregory Chanan, Peng Wu, and Soumith Chintala. PyTorch 2: Faster machine learning through dynamic Python bytecode transformation and graph compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. ACM, April 2024. doi: 10.1145/3620665.3640366.

[4] Resham Arya, Jaiteg Singh, and Ashok Kumar. A survey of multidisciplinary domains contributing to affective computing. *Computer Science Review*, 40:100399, 2021. ISSN 1574-0137. doi: 10.1016/j.cosrev.2021.100399.

[5] Tetsuya Asakawa, Riku Tsuneda, and Masaki Aono. Visual sentiment analysis multiplying deep learning and vision transformers. In *Proceedings of the MediaEval 2021 Workshop*, 12 2021.

[6] Neha Bhardwaj and Manish Dixit. A review: Facial expression detection with its techniques and application. *International Journal*

of Signal Processing, Image Processing and Pattern Recognition, 9: 149–158, 6 2016. ISSN 20054254. doi: 10.14257/ijsip.2016.9.6.13.

[7] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia*, pages 223–232. ACM, 10 2013. ISBN 9781450324045. doi: 10.1145/2502081.2502282.

[8] Cristina Bustos, Carles Civit, Brian Du, Albert Solé-Ribalta, and Agata Lapedriza. On the use of vision-language models for visual sentiment analysis: A study on clip. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

[9] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. doi: 10.48550/arXiv.1405.3531.

[10] Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv preprint arXiv:1410.8586*, 10 2014.

[11] Francois Chollet. Xception: Deep llearning with depthwise separable convolutions . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, Los Alamitos, CA, USA, Jul 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.195.

[12] Karina Cortinas-Lorenzo and Gerard Lacey. Toward explainable affective computing: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. ISSN 21622388. doi: 10.1109/TNNLS.2023.3270027.

[13] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on cComputer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[14] Edwin S. Dalmaijer, Camilla L. Nord, and Duncan E. Astle. Statistical power for cluster analysis. *BMC Bioinformatics*, 23, 12 2022. ISSN 14712105. doi: 10.1186/s12859-022-04675-1.

[15] DataReportal. Digital 2025: Global overview report. https://datareportal.com/reports/digital-2025-global-overview-report, February 2025. We Are Social; Kepios; Meltwater. Accessed: 2025-09-21.

[16] Renuka S. Deshmukh, Vandana Jagtap, and Shilpa Paygude. Facial emotion recognition system through machine learning approach. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 272–277, 2017. doi: 10.1109/ICCONS.2017.8250725.

[17] M Dewan, Mahbub Murshed, and Fuhua Lin. Engagement detection in online learning: A review. *Smart Learning Environments*, 6 (1):1–20, 2019.

[18] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426, 2015.

[19] Tuomas Eerola and Jonna K. Vuoskoski. A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, 39(1):18–49, 2011. doi: 10.1177/0305735610362821.

[20] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

[21] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

[22] Omar Elharrouss, Younes Akbari, Noor Almadeed, and Somaya Al-Maadeed. Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. *Computer Science Review*, 53:100645, August 2024. ISSN 1574-0137. doi: 10.1016/j.cosrev.2024.100645.

[23] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[24] Raúl Gómez et al. Icdar2017 robust reading challenge on coco-text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1435–1443. IEEE, 2017. ISBN 978-1-5386-3586-5. doi: 10.1109/ICDAR.2017.234.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learn-*

ing. MIT Press, 2016. http://www.deeplearningbook.org.

[26] Abhay Gupta, Vineeth Balasubramanian, Arjun D ' Cunha, and Kamal Awasthi. Daisee: Dataset for affective states in e-learning environments daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*, 2016. doi: 10.48550/arXiv.1609.01885.

[27] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Clustering validity checking methods: Part ii. *ACM SIGMOD Record*, 31(3):19–27, 2002. doi: 10.1145/507338.507342.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[29] Atis Hermanis, Ricards Cacurs, Krisjanis Nesenbergs, Modris Greitans, Emil Syundyukov, and Leo Selavo. Wearable sensor system for human biomechanics monitoring. In *EWSN '16: Proceedings of the 2016 International Conference on Embedded Wireless Systems and Networks*, pages 247–248, Graz, Austria, February 2016. ACM.

[30] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. *Information Fusion*, 49:69–78, 2019.

[31] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[32] J. L.Mazher Iqbal, M. Senthil Kumar, Geetishree Mishra, G. R. Asha, A. N. Saritha, A. Karthik, and N. Bonthu Kotaiah. Facial emotion recognition using geometrical features based deep learning techniques. *International Journal of Computers, Communications and Control*, 18, 2023. ISSN 18419844. doi: 10.15837/ijccc.2023.4.4644.

[33] Deepak Kumar Jain, Pourya Shamsolmoali, and Paramjit Sehdev. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120:69–74, 4 2019. ISSN 01678655. doi: 10.1016/j.patrec.2019.01.008.

[34] Shao Jie and Qian Yongsheng. Three convolutional neural network models for facial expression recognition in the wild. *Neurocomput-*

ing, 355:82–92, 05 2019. doi: 10.1016/j.neucom.2019.05.005.

[35] Arturas Kaklauskas, Edmundas Kazimieras Zavadskas, Valdas Pruskus, Andrejus Vlasenko, Lina Bartkiene, Rasa Paliskiene, Lina Zemeckyte, V Gerstein, Gintautas Dzemyda, and Gintautas Tamulevicius. Recommended biometric stress management system. *Expert Systems with Applications*, 38(11):14011–14025, 2011.

[36] Hang-Bong Kang. Affective content detection using HMMs. In *Proceedings of the Eleventh ACM International Conference on Multimedia*, MULTIMEDIA '03, page 259–262, New York, NY, USA, 2003. Association for Computing Machinery. ISBN 1581137222. doi: 10.1145/957013.957066.

[37] Rasa Karbauskaite, Leonidas Sakalauskas, and Gintautas Dzemyda. Kriging predictor for facial emotion recognition using numerical proximities of human emotions. *Informatica*, 31:249–275, 2020. ISSN 08684952. doi: 10.15388/20-INFOR419.

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN 0001-0782. doi: 10.1145/3065386.

[39] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[40] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2019-October, pages 10142–10151. Institute of Electrical and Electronics Engineers Inc., 10 2019. ISBN 9781728148038. doi: 10.1109/ICCV.2019.01024.

[41] Iulia Lefter, David D. Luxton, Alice Baird, Theodora Chaspari, Zakia Hammal, Marwa Mahmoud, and Albert Ali Salah. Affective computing for mental wellbeing: Challenges, opportunities, and promising synergies. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2023*. Institute of Electrical and Electronics Engineers Inc., 2023. ISBN 9798350327458. doi: 10.1109/ACIIW59127.2023.10388209.

[42] Liefa Liao, Shouluan Wu, Chao Song, and Jianglong Fu. Rs-xception: A lightweight network for facial expression recogni-

tion. *Electronics (Switzerland)*, 13, 8 2024. ISSN 20799292. doi: 10.3390/electronics13163217.

[43] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. doi: 10.48550/arXiv.1608.03983.

[44] Yutong Luo, Xinyue Zhong, Minchen Zeng, Jialan Xie, Shiyuan Wang, and Guangyuan Liu. Cglf-net: Image emotion recognition network by combining global self-attention features and local multiscale features. *IEEE Transactions on Multimedia*, 26:1894–1908, 2024. ISSN 19410077. doi: 10.1109/TMM.2023.3289762.

[45] Yutong Luo, Xinyue Zhong, Jialan Xie, and Guangyuan Liu. Cvrsf-net: Image emotion recognition by combining visual relationship features and scene features. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025. ISSN 2471285X. doi: 10.1109/TETCI.2025.3543300.

[46] Jana Machajdik and Allan Hanbury. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*, pages 83–92, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-933-6. doi: 10.1145/1873951.1873965.

[47] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. arXiv:1802.03426.

[48] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

[49] Joseph A. Mikels, Barbara L. Fredrickson, Gregory R. Larkin, Casey M. Lindberg, Sam J. Maglio, and Patricia A. Reuter-Lorenz. Emotional category data on images from the international affective picture system. *Behavior Research Methods*, 37:626–630, 11 2005. ISSN 1554-351X. doi: 10.3758/BF03192732.

[50] Saif M. Mohammad and Svetlana Kiritchenko. An annotated dataset of emotions evoked by art. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan, 2018.

[51] Modestas Motiejauskas and Gintautas Dzemyda. Efficientnet

convolutional neural network with gram matrices modules for predicting sadness emotion. *International Journal of Computers Communications & Control*, 19 (5):6697, 2024. doi: 10.15837/ijccc.2024.5.6697.

[52] Modestas Motiejauskas and Gintautas Dzemyda. The effective evaluation of emotions in the visual emotion images using convolutional neural networks. *IEEE Access*, 13:139174–139187, 2025. ISSN 21693536. doi: 10.1109/ACCESS.2025.3596484.

[53] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE Transactions on Affective Computing*, 5(2):101–111, 2014.

[54] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, page 443–449, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450339124. doi: 10.1145/2818346.2830593.

[55] T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of 12th International Conference on Pattern Recognition*, volume 1, pages 582–585 vol.1, 1994. doi: 10.1109/ICPR.1994.576366.

[56] Steve Pan, Jinsoo Lee, and Henry Tsai. Travel photos: Motivations, image dimensions, and affective qualities of places. *Tourism Management*, 40:59–69, 2014.

[57] W. Gerrod Parrott. *Emotions in Social Psychology: Essential Readings*. Psychology Press, 2001.

[58] Robert Plutchik. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4): 344–350, 2001.

[59] Ce Qi and Fei Su. Contrastive-center loss for deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2851–2855. IEEE, 9 2017. ISBN 978-1-5090-2175-8. doi: 10.1109/ICIP.2017.8296803.

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh,

Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

[61] Panda Rameswar, Zhang Jianming, Li Haoxiang, Lee Joon-Young, Lu Xin, and Roy-Chowdhury Amit K. Contemplating visual emotions: Understanding and overcoming dataset bias. In Ferrari Vittorio, Hebert Martial, Sminchisescu Cristian, and Weiss Yair, editors, *Computer Vision – ECCV 2018*, pages 594–612, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01216-8.

[62] Andrew G. Reece and Christopher M. Danforth. Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6, 12 2017. ISSN 21931127. doi: 10.1140/epjds/s13688-017-0110-z.

[63] I. Michael Revina and W. R.Sam Emmanuel. A survey on human face expression recognition techniques. *Journal of King Saud University - Computer and Information Sciences*, 2018. ISSN 22131248. doi: 10.1016/j.jksuci.2018.09.002.

[64] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018.

[65] Xue Rui. A convolutional neural networks based approach for clustering of emotional elements in art design. *PeerJ Computer Science*, 9, 2023. ISSN 23765992. doi: 10.7717/peerj-cs.1548.

[66] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 12 2015. ISSN 15731405. doi: 10.1007/s11263-015-0816-y.

[67] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161, 1980.

[68] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, 2018. doi: 10.1109/CVPR.2018.00474.

[69] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sut-

skever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

[70] Ben A. Steward, Paige Mewton, Romina Palermo, and Amy Dawel. Interactions between faces and visual context in emotion perception: A meta-analysis. *Psychonomic Bulletin & Review*, 32(5):1987–2003, 2025. doi: 10.3758/s13423-025-02678-6.

[71] Jianshan Sun, Qing Zhang, Kun Yuan, Yuanchun Jiang, and Xinran Chen. A supervised contrastive learning-based model for image emotion classification. *World Wide Web*, 27, 5 2024. ISSN 15731413. doi: 10.1007/s11280-024-01260-9.

[72] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[73] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[74] Mingxing Tan and Quoc Le. EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR, 2021.

[75] Juan Terven, Diana Margarita Cordova-Esparza, Julio Alejandro Romero-González, Alfonso Ramírez-Pedraza, and E. A. Chávez-Urbiola. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58, 7 2025. ISSN 15737462. doi: 10.1007/s10462-025-11198-7.

[76] Robert E. Thayer. *The Biopsychology of Mood and Arousal*. Oxford University Press, 1990.

[77] Mesut Ulu and Yusuf Sait Türkan. Cluster analysis and comparative study of different clustering performance and validity indices. In Numan M. Durakbasa and M. Güneş Gençyılmaz, editors, *Industrial Engineering in the Industry 4.0 Era*, Lecture Notes in Mechanical Engineering, pages 33–45. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-53991-6. doi: 10.1007/978-3-031-53991-6_3.

[78] Shangfei Wang, Menghua He, Zhen Gao, Shan He, and Qiang

Ji. Emotion recognition from thermal infrared images using deep boltzmann machine. *Frontiers of Computer Science*, 8(4):609–618, 2014.

[79] Matthijs J. Warrens and Hanneke van der Hoef. Understanding the adjusted rand index and other partition comparison indices based on counting object pairs. *Journal of Classification*, 39:487–509, 11 2022. ISSN 14321343. doi: 10.1007/s00357-022-09413-z.

[80] Markus Wegmann, Domenique Zipperling, Jonas Hillenbrand, and Jürgen Fleischer. A review of systematic selection of clustering algorithms and their evaluation. *ArXiv*, abs/2106.12792, 2021.

[81] Liwen Xu, Zhengtao Wang, Bin Wu, and Simon Lui. Mdan: Multi-level dependent attention network for visual emotion analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9479–9488, 2022.

[82] Qinfu Xu, Yiwei Wei, Shaozu Yuan, Jie Wu, Leiquan Wang, and Chunlei Wu. Learning emotional prompt features with multiple views for visual emotion analysis. *Information Fusion*, 108, 8 2024. ISSN 15662535. doi: 10.1016/j.inffus.2024.102366.

[83] Qinfu Xu, Shaozu Yuan, Yiwei Wei, Jie Wu, Leiquan Wang, and Chunlei Wu. Multiple feature refining network for visual emotion distribution learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8924–8932, 2025.

[84] Wanni Xu, You Lei Fu, and Dongmei Zhu. Resnet and its application to medical image processing: Research progress and challenges. *Computer Methods and Programs in Biomedicine*, 240, 10 2023. ISSN 18727565. doi: 10.1016/j.cmpb.2023.107660.

[85] Hansen Yang, Yangyu Fan, Guoyun Lv, Shiya Liu, and Zhe Guo. Exploiting emotional concepts for image emotion recognition. *The Visual Computer*, 39(5):2177–2190, 2023. doi: 10.1007/s00371-022-02472-8.

[86] Jiajun Yang, Lianggui Tang, Zhuo Chen, Xiuling Zhu, and Xuan Lai. Facial emotion recognition based on optimized xception training. In *ACM International Conference Proceeding Series*, pages 12–18. Association for Computing Machinery, 12 2024. ISBN 9798400717529. doi: 10.1145/3697355.3697358.

[87] Jingyuan Yang, Jie Li, Leida Li, Xiumei Wang, and Xinbo Gao. A circular-structured representation for visual emotion distribution

learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4237–4246, 2021.

[88] Jingyuan Yang, Jie Li, Xiumei Wang, Yuxuan Ding, and Xinbo Gao. Stimuli-aware visual emotion analysis. *IEEE Transactions on Image Processing*, 30:7432–7445, 2021. doi: 10.1109/TIP.2021.3106813.

[89] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20326–20337, 2023. doi: 10.1109/ICCV51070.2023.01864.

[90] Jin Ye, Xiaojiang Peng, Yu Qiao, Hao Xing, Junli Li, and Rongrong Ji. Visual-textual sentiment analysis in product reviews. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 869–873. IEEE, 2019.

[91] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Building a large scale dataset for image emotion recognition: The fine print and the benchmark. *arXiv preprint arXiv:1605.02677*, May 2016. doi: 10.48550/arXiv.1605.02677.

[92] Tong Yu and Hong Zhu. Hyper-parameter optimization: A review of algorithms and applications. *arXiv preprint arXiv:2003.05689*, 2020.

[93] Rajamanickam Yuvaraj, Rakshit Mittal, A. Amalin Prince, and Jun Song Huang. Affective computing for learning in education: A systematic review and bibliometric analysis. *Education Sciences*, 15, 1 2025. ISSN 22277102. doi: 10.3390/educsci15010065.

[94] Haimin Zhang and Min Xu. Weakly supervised emotion intensity prediction for recognition of emotions in images. *IEEE Transactions on Multimedia*, 23:2033–2044, 2021. ISSN 19410077. doi: 10.1109/TMM.2020.3007352.

[95] Haimin Zhang and Min Xu. Multiscale emotion representation learning for affective image recognition. *IEEE Transactions on Multimedia*, 25:2203–2212, 2023. ISSN 19410077. doi: 10.1109/TMM.2022.3144804.

[96] Hao Zhang, Yanan Liu, Dan Xu, Kangjian He, Guoqing Peng, Yingying Yue, and Ruhan Liu. Learning multi-level representations for image emotion recognition in the deep convolutional network. In *Proceedings of SPIE, the International Society for Optical*

Engineering (Vol. 12083), page 91. SPIE-Intl Soc Optical Eng, 2 2022. ISBN 9781510650428. doi: 10.1117/12.2623414.

[97] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[98] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. Attention based fully convolutional network for speech emotion recognition. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1771–1775. IEEE, 2018.

[99] Guangzhe Zhao, Hanting Yang, Bing Tu, and Lei Zhang. A survey on image emotion recognition. *Journal of Information Processing Systems*, 17(6):1138–1156, 2021. ISSN 1976-913X. doi: 10.3745/JIPS.01.0082.

[100] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM International Conference on Multimedia (MM '14)*, pages 47–56, New York, NY, USA, November 2014. ACM. ISBN 978-1-4503-3063-3. doi: 10.1145/2647868.2654930.

[101] Sicheng Zhao, Guiguang Ding, Qingming Huang, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: A comprehensive survey. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5534–5541. International Joint Conferences on Artificial Intelligence Organization, July 2018. doi: 10.24963/ijcai.2018/780.

[102] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Björn W. Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6729–6751, October 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3094362.

SANTRAUKA (SUMMARY IN LITHUANIAN)

S.1 ĮVADAS

Emocijos atlieka svarbų vaidmenį žmogaus komunikacijoje, nes daro įtaką pažinimo procesams, sprendimų priėmimui ir socialinei sąveikai. Spartus skaitmeninių technologijų vystymasis ir vaizdinės terpės dominavimas skatina poreikį geriau suprasti emocinį vaizdų turinį. Ši problema tampa ypač svarbi emocijų kompiuterinės analizės (angl. affective computing) srityje – tarpdisciplininėje mokslinių tyrimų kryptyje, kurios tikslas yra kurti sistemas, gebančias atpažinti, interpretuoti ir suvokti žmogaus emocijas. Šiame kontekste vizualinių emocijų atpažinimas apibrėžiamas kaip automatinis emocijų atpažinimas vaizduose.

Emocijų kompiuterinės analizės tyrimuose dažnai taikoma Mikels et al. [49] pasiūlyta taksonomija, apimanti aštuonias diskrečias emocijų kategorijas ir plačiai naudojama vizualinių emocijų duomenų rinkiniuose. Šioje disertacijoje remiamasi būtent Mikels taksonomija, nes ši atitinka eksperimentuose naudojamų duomenų rinkinių emocijų žymėjimą.

S.1.1 Tyrimo problema

Vizualinių emocijų atpažinimo srityje pagrindiniai iššūkiai yra emocinės išraiškos atotrūkis, emocijų suvokimo subjektyvumas ir vizualinių emocijų duomenų žymėjimo sudėtingumas [102]. Emocinės išraiškos atotrūkis nusako neatitikimą tarp fizinių vaizdo požymių ir emocinės būsenos, kurią vaizdas sukelia stebėtojui, todėl emocinė interpretacija dažnai priklauso ne tik nuo atskirų požymių, bet ir nuo viso vaizdo kompozicijos bei konteksto. Skirtingi stebėtojai tą patį vaizdą taip pat gali interpretuoti nevienodai [99].

Išlieka aktuali problema, kaip konvoliucinių neuroninių tinklų (CNN) pagrindu sukurti modelį, gebantį veiksmingai integruoti semantinius ir stilistinius požymius, pagerinti emocijų klasių atskiriamumą požymių erdvėje ir užtikrinti patikimesnį emocijų klasifikavimą bendro pobūdžio

vaizduose. Taip pat trūksta metodų, leidžiančių įvertinti išmoktų išraiškų struktūrą ir vidinį modelio prognozių nuoseklumą.

S.1.2 Aktualumas

Didėjantis vizualinių duomenų kiekis skaitmeninėje erdvėje [15] rodo automatizuotų metodų, gebančių interpretuoti emocinį vaizdų turinį, poreikį. Šioje disertacijoje siūlomas CNN pagrindu sukurtas modelis ir mokymo strategija, skirti semantiniams ir stilistiniams požymiams gauti bei emocijų klasių atskiriamumui požymių erdvėje gerinti. Modelyje integruojamos Gramo matricos pagrindu formuojamos požymių išraiškos, taikomas bendras optimizavimas su kontrastinių centrų nuostolio funkcija, o papildomai išmoktoms išraiškoms vertinti siūlomas top-2 priešingų sentimentų nuoseklumo matas.

S.1.3 Tyrimų sritis

Ši disertacija plėtojama kuriant ir vertinant giliojo mokymosi modelius, skirtus geriau fiksuoti sudėtingus vizualinius veiksnius, turinčius įtakos emocijų suvokimui.Dėmesys skiriamas požymių išraiškai tobulinti ir emocijų klasių atskiriamumui didinti, taikant optimizuotas neuroninių tinklų architektūras ir nuostolių funkcijas.

S.1.4 Disertacijos objektas

Disertacijos objektas – emocijas perteikiantys vaizdai ir jų analizei skirti konvoliuciniai neuroniniai tinklai.

S.1.5 Disertacijos tikslas

Disertacijos tikslas – CNN pagrindu sukurti ir įvertinti sukurtą modelį vizualinėms emocijos atpažinti, gebantį užfiksuoti su emocijomis susijusias vizualines savybes ir pagerinti emocijų klasių atskiriamumą.

S.1.6 Disertacijos uždaviniai

Siekiant įgyvendinti iškeltą tikslą, suformuluoti šie uždaviniai:

- Atlikti literatūros analizę, siekiant nustatyti pagrindinius vizualinių emocijų atpažinimo iššūkius, taikant giliojo mokymosi metodus.
- Atlikti pasirinktų CNN architektūrų lyginamąją empirinę analizę, siekiant nustatyti tinkamiausią pagrindinį tinklą vizualinėms emocijoms atpažinti.
- Sukurti CNN pagrindu paremtą modelį vizualinėms emocijoms atpažinti bendro pobūdžio vaizduose, didinant požymių išraiškų kokybę ir mažinant emocinės išraiškos atotrūkį.
- Suprojektuoti ir įvertinti CNN pagrindu sukurtą modelį, integruojant kontrastinių centrų nuostolio funkciją, siekiant pagerinti emocijų klasių atskiriamumą.
- Išanalizuoti išmoktos požymių erdvės struktūrą, taikant dimensijų mažinimą, klasterizavimo metrikas ir top-2 priešingų sentimentų nuoseklumo matą, siekiant įvertinti emocijų klasių atskiriamumo ir išraiškos struktūros pokyčius.

S.1.7 Tyrimo metodai

Disertacijoje taikyta literatūros analizė, CNN pagrindu sukurto modelio kūrimas bei eksperimentiniai ir analitiniai metodai veiksmingam vizualinių emocijų atpažinimui. Tyrimo metodai pateikiami tokia tvarka:

- Teoriniai metodai:
  Literatūros apžvalga, analizė ir apibendrinimas buvo taikomi siekiant apibrėžti tyrimo problemą, išskirti emocinės išraiškos atotrūkį kaip svarbų VER iššūkį ir pagrįsti siūlomo metodo naujumą.

- Eksperimentiniai metodai:
  Sukurtas CNN pagrindu paremtas modelis. Tai apėmė Gramo matricos modulių integravimą, kontrastinių centrų nuostolio funkcijos įgyvendinimą bei modelio mokymo ir vertinimo procedūras. Taip pat buvo taikomi duomenų paruošimo ir augmentacijos metodai.

- Analitiniai metodai:
  Modelio veikimas vertintas naudojant klasifikavimo metrikas (tikslumą, preciziškumą, jautrumą, $F1$ rodiklį) ir mūsų siūlomą top-2 priešingų sentimentų vertinimą. Išmoktos požymių erdvės struktūra analizuota taikant dimensijų mažinimo, klasterizavimo ir klasterių kokybės metrikas (angl. adjusted Rand index ARI, normalized mutual information NMI, ambiguous sample ratio ASR), siekiant įvertinti klasių atskiriamumą ir požymių išraiškos kokybę. Atlikta lyginamoji analizė su baziniais modeliais.

S.1.8 Mokslinis naujumas

Disertacijoje pristatomi keli mokslinio naujumo aspektai vizualinių emocijų analizės srityje. Pagrindinis šios disertacijos indėlis – CNN pagrindu sukurto modelio patikimesnei vizualinių emocijų klasifikacijai kūrimas ir įvertinimas. Šis indėlis grindžiamas modelio plėtojimu integruojant papildomas požymių išraiškas emocinės išraiškos atotrūkiui mažinti. Be to, pasiūlytas kontrastinių centrų nuostolio komponentas, skirtas geriau atskirti emocijų klases apmokant modelį.

Disertacijoje išskiriami šie mokslinio naujumo aspektai:

- Pasiūlyta nauja mokymo optimizavimo strategija, kurioje kontrastinių centrų nuostolio funkcija optimizuojama kartu su praretinta kategorine kryžminės entropijos nuostolių funkcija (angl. sparse categorical cross-entropy), siekiant pagerinti vizualinių emocijų klasių atskiriamumą.

- Pasiūlyta Gramo matricos modulių integracija į CNN architektūrą, siekiant papildyti modelį stilistiniais požymiais.

- Pasiūlyta metodika išmoktos požymių erdvės struktūrai analizuoti ir kontrastinių centrų nuostolio integravimo veiksmingumui mokymo procese įvertinti.

- Pasiūlytas top-2 priešingų sentimentų nuoseklumo vertinimo matas modelio prognozių vidiniam nuoseklumui vertinti, papildantis įprastinį vertinimą be tikrųjų žymėjimų poreikio.

S.1.9 Tyrimo praktinė vertė

Šios disertacijos praktinė vertė slypi siūlomame CNN pagrindu sukurtame modelyje vizualinėms emocijoms atpažinti. Gramo matricos modulių integracija suteikia veiksmingą būdą fiksuoti stilistinius vaizdų požymius, tokius kaip spalva, tekstūra ir pasikartojantys raštai, kurie yra svarbūs emocinei išraiškai. Tai leidžia užtikrinti patikimesnę emocijų klasifikaciją įvairiuose vaizdų kontekstuose.

Kontrastinių centrų nuostolio įtraukimas į mokymo procesą sustiprina emocijų kategorijų atskiriamumą požymių erdvėje, todėl prognozės tampa patikimesnės ir sumažėja painiava tarp vizualiai panašių klasių. Šie patobulinimai gali būti tiesiogiai taikomi emocijų analizės sistemose, įskaitant turinio rekomendavimo sistemas, psichikos sveikatos stebėseną, žmogaus ir kompiuterio sąveiką bei afektinės daugialypės terpės paiešką.

Be to, šioje disertacijoje sukurtas eksperimentinis karkasas, apimantis duomenų paruošimo grandines, išraiškos požymių vizualizaciją, klasterizavimo ir klasterių kokybės vertinimo metrikas bei priešingų sentimentų matą, sudaro metodinį pagrindą tolesniems emocijų kompiuterinės analizės tyrimams.

S.1.10 Ginamieji teiginiai

- Gramo matricos modulių integravimas į EfficientNetV2S modelį leidžia gauti papildomus žemo lygmens požymius, tokius kaip

spalva, tekstūra ir pasikartojantys raštai. Derinant juos su semantiniais CNN požymiais, gaunama išsamesnė išraiška, padedanti geriau atskleisti emocinį vaizdo turinį.

- Siūlomas kontrastinių centrų nuostolis padidina emocijų klasių atskiriamumą, suspausdamas tos pačios klasės išraiškos požymius ir didindamas atstumus tarp skirtingų klasių, todėl klasifikavimo rezultatai pagerėja, palyginti su mokymu taikant tik kryžminės entropijos nuostolį.

- Išmoktos požymių erdvės analizė, taikant dimensijų mažinimą, klasterizavimo metrikas ir priešingų sentimentų nuoseklumo matą, rodo, kad siūlomas modelis išmoksta nuoseklesnes ir geriau struktūruotas emocijų išraiškas nei bazinės CNN architektūros.

- Top-2 priešingų sentimentų rodiklis papildo tikslumu paremtą vertinimą be tikrųjų klasių žymėjimo vertinimo poreikio. Šis rodiklis nusako, kaip dažnai dvi didžiausią tikimybę turinčios modelio prognozės priklauso priešingoms sentimentų grupėms. Kontrastinių centrų nuostolio integravimas sumažina tokių priešingų atvejų skaičių, palyginti su baziniu modeliu.

S.2 TYRIMO REZULTATŲ APROBACIJA

Pagrindiniai šios disertacijos rezultatai publikuoti recenzuojamuose moksliniuose žurnaluose ir pristatyti tarptautinėse bei nacionalinėse konferencijose.

Publikacijos Clarivate Web of Science žurnaluose

[A1] Modestas Motiejauskas, Gintautas Dzemyda (2024). *EfficientNet Convolutional Neural Network with Gram Matrices Modules for Predicting Sadness Emotion*. International Journal of Computers Communications & Control, 19(5), art. no. 6697.
https://doi.org/10.15837/ijccc.2024.5.6697

[A2] Modestas Motiejauskas, Gintautas Dzemyda (2025). *The Effective Evaluation of Emotions in the Visual Emotion Images Using Convolutional Neural Networks*. IEEE Access, 13, 139174–139187. https://doi.org/10.1109/ACCESS.2025.3596484

Publikacijos tarptautinių konferencijų recenzuojamuose leidiniuose

[B1] Modestas Motiejauskas, Gintautas Dzemyda (2024). *Evaluation of Emotions in Artworks Using EfficientNet Convolutional Network Integrating the Gram Matrix Modules*. In: 2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), Xiamen, China, 882–887. https://doi.org/10.1109/ICAIRC64177.2024.10900186

Pristatymai tarptautinėse konferencijose

[C1] Modestas Motiejauskas, Gintautas Dzemyda (2023). *Optimization of EfficientNetV2 Models for Predicting Sadness Emotion*. Numerical Computations: Theory and Algorithms (NUMTA-2023), Calabria, Italy, June 14 - 20, 2023.

[C2] Modestas Motiejauskas, Gintautas Dzemyda (2024). *Evaluation of Emotions in Artworks Using EfficientNet Network Integrating the Gram Matrix Modules*. 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC), Xiamen, China, December 27-29, 2024.

Pristatymai nacionalinėse konferencijose

[D1] Modestas Motiejauskas, Gintautas Dzemyda (2022). *On recognizing emotion of sadness in images of a general nature using CNN*. Data Analysis Methods for Software Systems, Druskininkai, Lietuva, December 01-03, 2022.

S.3 LITERATŪROS APŽVALGA

Šiame skyriuje apžvelgiamos pagrindinės tyrimų kryptys, pasiekti rezultatai ir emocijų atpažinimo vaizduose srityje kylančios problemos.

S.3.1 Vizualinė emocijų analizė

Emocijų kompiuterinė analizė – tai tarpdisciplininė mokslo sritis, kurios tikslas yra suteikti kompiuterinėms sistemoms gebėjimą atpažinti, interpretuoti ir modeliuoti žmogaus emocijas, analizuojant multimodalius duomenis, tokius kaip tekstas, kalba, fiziologiniai signalai (pvz., elektroencefalograma, elektrokardiograma) ir vizualiniai dirgikliai [4], [53]. Šios srities taikymo kryptys apima žmogaus ir kompiuterio sąveiką, psichikos sveikatą [41] bei švietimą [93]. Platesniame kontekste vizualinė emocijų analizė (angl. visual emotion analysis VEA) išskiriama kaip specializuota sritis, nagrinėjanti emocines reakcijas, kurias sukelia vaizdai.

Kituose tyrimuose dažnai vartojami susiję terminai, tokie kaip emocinę išraišką perteikiančių vaizdų turinio analizė (angl. affective image content analysis AICA) [101], [102]. AICA tiria ryšį tarp vaizdo turinio, emocijų ir susijusių disciplinų (pvz., psichologijos, sociologijos), taip pat praktinius taikymus realiose situacijose. O VEA [89] orientuojasi siauriau - į skaičiavimo metodus, taikomus vaizdiniams požymiams gauti ir emocijoms klasifikuoti. Ši sritis akcentuoja algoritminius emocijų klasifikavimo metodus, o ne tarpdisciplininius aspektus.

AICA nagrinėja, kaip vaizdai susiję su emocijomis ar jų būsenomis įvairiose disciplinose, o VEA pirmenybę teikia techniniams sprendimams – požymiams gauti, mašininio mokymosi modeliams ir duomenų rinkiniams kurti – siekiant kiekybiškai įvertinti emocines reakcijas į vizualinius duomenis. Šioje disertacijoje emocinę išraišką perteikiančių vaizdų analizės sritis apribojama vizualine emocijų analize, konkrečiai sprendžiant emocijų atpažinimo iš statinių vaizdų klasifikavimo uždavinį.

S.1 paveiksle pavaizduoti pagrindiniai VEA etapai. Kiekvienas iš jų

S.1 pav.: Kiekvienas etapas atitinka pagrindinį VEA proceso komponentą – nuo požymių gavimo iki emocijų klasifikavimo, išryškinant duomenų srautą ir tarpusavio priklausomybes emocijų analizės procese.

detaliau aprašomas tolesniuose skyriuose.

Emocijų atpažinimas yra tiriamas įvairiomis susijusiomis formomis, įskaitant tekstą, kalbą, lingvistiką, muziką, garsą, veido išraiškų analizę, vaizdo įrašus, fiziologinius signalus ir multimodalius duomenis. VEA orientuojama į emocinio turinio analizę vaizduose, vizualinių požymių gavimą ir interpretavimą. Emocijų atpažinimo problemą galima suskirstyti į tris etapus: žmogaus anotavimą, vizualinių požymių gavimą ir mokymąsi, kurio metu gauti požymiai susiejami su suvokiama emocija. Literatūroje dažnai pabrėžiama, kad pagrindinis skirtumas tarp AICA ir tipinių kompiuterinės regos uždavinių yra vadinamasis emocinės išraiškos atotrūkis. Jis gali būti iliustruojamas taip: fizinis rožės objektas šviesiame fone gali kelti teigiamą emociją, bet tas pats objektas

tamsiame, niūriame fone gali būti suvokiamas kaip keliantis neigiamą emociją.

VEA srityje emocijos paprastai išreiškiamos naudojant dvi pagrindines taksonomijų rūšis: kategorines emocijų būsenas (CES) ir dimensines emocijų erdves (DES). Šios taksonomijos apibrėžia emocijų kategorijas arba dimensijas, naudojamas giliojo mokymosi modeliuose, bei lemia modelių mokymo ir vertinimo principus. Dauguma viešai prieinamų vizualinių emocijų duomenų rinkinių paremti viena iš šių taksonominių struktūrų, kurios yra esminės tiek duomenų rinkiniams kurti, tiek modeliams vystyti.

S.4 NAUJAS MODELIS VIZUALINEI EMOCIJŲ ANALIZEI

Šiame skyrelyje pateikiama metodikos sistema, taikanti Gramo matricų modulius vizualinei emocijų analizei gerinti. Problema formuluojama kaip vizualinių emocijų klasifikavimo uždavinys. Be to, šiame skyrelyje siūlomi modelio patobulinimai emocijų atpažinimo patikimumui ir atsparumui didinti. Modelio elgsenai tirti ir analizuoti taikomos vizualizacijos technikos, leidžiančios kokybiškai įvertinti išmoktų požymių tarpusavio ryšius bei nustatyti galimas spragas sukurtame vizualinių emocijų atpažinimo modelyje.

Tyrimo metodika ir jos dalys publikuoti recenzuojamuose moksliniuose straipsniuose [A1, A2, B1].

S.4.1 Tyrimo metodika ir siūlomas metodas

Šiame skyrelyje pristatoma metodika vizualinių emocijų atpažinimo (VER) tikslumui gerinti. Remiantis literatūros apžvalgos (S.3 skyrelis) išvadomis, kuriose išryškinta semantinės abstrakcijos ir žemo lygmens suvokimo požymių balansavimo problema, šiame skyrelyje aprašomi sprendimai, eksperimentinė aplinka ir siūlomi modelio patobulinimai, skirti šioms spragoms spręsti.

Mūsų siūlomas metodas grindžiamas ankstesniu Zhang et al. [96] darbu, kuris parodė, kad negiliųjų (angl. shallow) sluoksnių Gramo mat-

ricos veiksmingai fiksuoja tekstūrinius ir spalvinius požymius, svarbius sentimentams klasifikuoti. Kitaip nei ankstesniame darbe, šiame tyrime įvedami Gramo matricos moduliai, atliekantys sluoksniui būdingą dimensijų mažinimą, leidžiantį gauti kompaktiškas požymių išraiškas skirtinguose CNN gylio lygiuose užtikrinant lankstų jų integravimą.

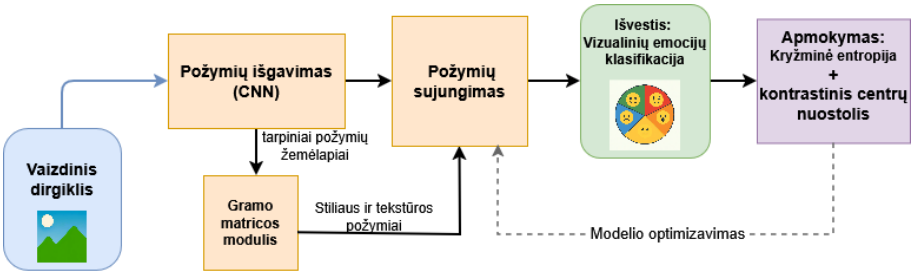
Papildomai taikomos vizualizacijos technikos, siekiant analizuoti modelio elgseną ir kokybiškai įvertinti požymių išraiškas skirtingose emocijų kategorijose. Šiame skyrelyje pristatoma metodika vizualinių emocijų atpažinimo (VER) tikslumui gerinti. Remiantis literatūros apžvalgos (S.3 skyrelis) išvadomis, kuriose išryškinta semantinės abstrakcijos ir žemo lygmens suvokimo požymių balansavimo problema, šiame skyrelyje aprašomi sprendimai, eksperimentinė aplinka ir siūlomi modelio patobulinimai, skirti šioms spragoms spręsti.

Mūsų siūlomas metodas grindžiamas ankstesniu Zhang et al. [96] darbu, kuris parodė, kad negiliųjų (angl. shallow) sluoksnių Gramo matricos veiksmingai fiksuoja tekstūrinius ir spalvinius požymius, svarbius sentimentams klasifikuoti. Kitaip nei ankstesniame darbe, šiame tyrime įvedami Gramo matricos moduliai, atliekantys sluoksniui būdingą dimensijų mažinimą, leidžiantį gauti kompaktiškas požymių išraiškas skirtinguose CNN gylio lygiuose užtikrinant lankstų jų integravimą.

Papildomai taikomos vizualizacijos technikos, siekiant analizuoti modelio elgseną ir kokybiškai įvertinti požymių išraiškas skirtingose emocijų kategorijose.

Siūlomi pagalbiniai išplečiami Gramo matricos moduliai, kurie išsišakoja iš pagrindinio CNN ir kuriais gauti požymiai sujungiami galutinėje tinklo struktūroje. Pagrindiniai šio darbo indėliai, palyginti su Zhang et al. [96], yra šie:

- Sukurta CNN pagrįsta sistema su išplečiamais Gramo matricos moduliais vizualinėms emocijos klasifikuoti.
- Pasiūlyti Gramo matricos moduliai, atliekantys sluoksniui būdingą dimensijų mažinimą, leidžiantį gauti suspaustas išraiškas skirtingose CNN dalyse.
- Atlikta vidinių požymių išraiškų vizualizacija ir kokybinė analizė.



S.2 pav.: Siūloma CNN pagrindu sukurta vizualinių emocijų atpažinimo sistema.

S.2 paveiksle pavaizduota siūloma giliojo mokymosi pagrindu sukurta vizualinių emocijų atpažinimo sistema. Įvestis yra vaizdas, perteikiantis tam tikrą emociją. Požymiams gauti naudojamas EfficientNetV2 konvoliucinis neuroninis tinklas. Tarpiniai (angl. intermediate) tinklo požymių atvaizdžiai (angl. feature maps) integruojami į atitinkamus Gramo matricos modulius. Gramo moduliai gauna spalvos, tekstūros ir stiliaus išraiškas, kurios atitinka žemo lygmens požymius, svarbius vizualinėms emocijoms nustatyti.

Tarpinių požymių žemėlapių gavimas iš pagrindinio tinklo gali būti apibrėžtas kaip pasirinktų sluoksnių išvesčių paėmimas. Šie požymių žemėlapiai atspindi išmoktas vaizdines struktūras skirtinguose abstrakcijos lygmenyse, todėl pagrindinis tinklas dažnai laikomas požymių gavimo moduliu. Formaliai gautus požymių žemėlapius apibrėžiame, kaip nurodyta toliau.

Tegu CNN (atraminis) tinklas yra B, sudarytas iš sluoksnių $\{l_1, \ldots, l_L\}$. Apibrėžkime indeksų aibę $S \subseteq \{1, \ldots, L\}$ sluoksnių, iš kurių gaunami požymių atvaizdžiai, ir tegu $v = |S|$ žymi pasirinktų sluoksnių skaičių. Kiekvienam pasirinktam sluoksniui $l_i \in S$ tegu

$$F_{l_i} \in \mathbb{R}^{C_i \times H_i \times W_i} \tag{S.1}$$

žymi atitinkamą gautą požymių žemėlapį, čia C_i, H_i ir W_i atitinkamai žymi kanalų skaičių, erdvinį aukštį ir plotį sluoksnio l_i požymių žemėlapyje.

Kiekvienas požymių žemėlapis F_{l_i} perduodamas apmokomam Gra-

mo matricos moduliui $g_i(\cdot)$, kuris suformuoja požymių vektorių:

$$z_i = g_i(F_{l_i}), \qquad z_i \in \mathbb{R}^{d_i}. \tag{S.2}$$

Čia d_i yra atitinkamo Gramo matricos modulio požymių dimensija, apibrėžiama taip:

$$d_i = \lfloor C_i/2 \rfloor, \tag{S.3}$$

o z_i yra sluoksniui l_i atitinkančio Gramo matricos modulio išvesties požymių vektorius.

Gramo matricos modulių skaičius yra v. Visų Gramo matricos modulių išvestys sujungiamos į vieną požymių vektorių:

$$z = [z_1; z_2; \ldots; z_v] \in \mathbb{R}^D, \qquad D = \sum_{i=1}^{v} d_i. \tag{S.4}$$

Tegu $h \in \mathbb{R}^{C_*}$ žymi globalų požymių vektorių, gaunamą iš pagrindinio tinklo po globaliojo vidurkinimo, o C_* yra pagrindinio tinklo išvesties dimensija.

Kad pagrindinio tinklo išvestis būtų suderinama su sujungta Gramo matricos modulių išvestimi, apibrėžiama išmokstama projekcija (čia įgyvendinama visiškai sujungtu sluoksniu):

$$P : \mathbb{R}^{C_*} \to \mathbb{R}^D, \tag{S.5}$$

o projektuotas pagrindinio tinklo vektorius apibrėžiamas taip:

$$p = P(h), \qquad p \in \mathbb{R}^D. \tag{S.6}$$

Abi išraiškos sujungiamos atliekant elementinę sudėtį:

$$u = p + z, \qquad u \in \mathbb{R}^D, \tag{S.7}$$

čia u yra sujungtas priešpaskutinio sluoksnio požymių vektorius.

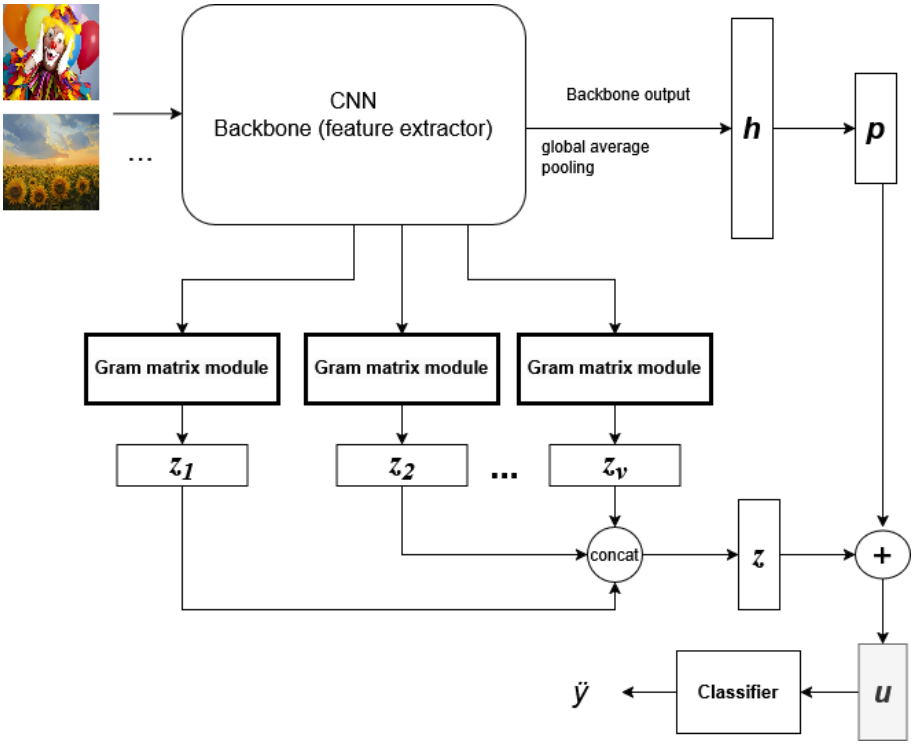
Galiausiai klasifikatorius susietą išraišką u paverčia išvesties prog-

nozės vektoriumi:

$$\hat{y} = \text{Klasifikatorius}(u), \qquad \hat{y} \in \mathbb{R}^N, \tag{S.8}$$

čia N yra emocijų klasių skaičius, o $\hat{y}$ žymi modelio išvesties logitų vektorių. Klasių tikimybės yra gaunamos logitų vektoriui $\hat{y}$ pritaikius softmax funkciją.

Pasirinktų pagrindinio tinklo požymių žemėlapio kanalų dimensijos yra $C_1 = 48$, $C_2 = 64$ ir $C_3 = 160$. Taikant aprašytą mažinimo taisyklę $d_i = \lfloor C_i/2 \rfloor$, gauname $(d_1, d_2, d_3) = (24, 32, 80)$, o sujungto visų Gramo matricos modulių išvesties vektoriaus, kai $v = 3$, dimensija yra $D = 136$.



S.3 pav.: Bendroji siūlomo modelio schema.

S.3 paveiksle h žymi pagrindinio tinklo išvestį po globaliojo vidurkinimo, p – projektuotą (suspaustą) pagrindinio tinklo požymių vektorių, $z_1, \ldots, z_v$ – atitinkamų Gramo matricos modulių išvestis, z – sujungtą visų Gramo matricos modulių išvesties vektorių, u – susietą priešpaskutinio sluoksnio požymių vektorių, o $\hat{y}$ - galutinę klasifikatoriaus išvestį.

Gramo matrica yra kvadratinės formos ir išreiškiama $C \times C$ matrica. Ji išskleidžiama į vienmatį vektorių, sudarytą iš $C \times C$ elementų, kuris toliau suspaudžiamas kitu sluoksniu iki $C/2$ vienetų. Paskui taikoma SiLU aktyvacijos funkcija [21] ir normalizavimas. Kita modulio šaka susideda iš 1×1 konvoliucijos operacijos, tada formuojamas rezultatas iš $C/2$ požymių. Kiekvienam iš jų apskaičiuojama vidutinė reikšmė per visas $H \times W$ padėtis, taip suformuojant $C/2$ ilgio vektorių. Galutinė Gramo matricos modulio išvestis gaunama sujungiant abi šakas elementinės sudėties būdu.

Gramo matrica $G \in \mathbb{R}^{C \times C}$ gali būti užrašyta taip:

$$G = FF^T, \qquad F \in \mathbb{R}^{C \times HW}, \quad F^T \in \mathbb{R}^{HW \times C}. \tag{S.9}$$

čia C žymi kanalų skaičių, o H ir W – požymių atvaizdžio aukštį ir plotį. Nors Gramo matrica apskaičiuojama iš vieno sluoksnio požymių atvaizdžio, ji apibendrina informaciją per visas erdvines pozicijas ir suteikia sluoksnio struktūrinį apibendrinimą stilistine prasme.

Gilesnių sluoksnių Gramo matricų reikšmės gali būti didesnės ir lemti skaitinį nestabilumą. Todėl taikoma normalizacija:

$$G = \frac{FF^T}{H \cdot W}. \tag{S.10}$$

Tokiu būdu sumažinamas reikšmių intervalas ir užtikrinamas stabilesnis tinklo apmokymas.

S.4 paveiksle pateikiamas pavyzdys, kaip Gramo matricos pagrindu sudarytos išraiškos koduoja vizualinę informaciją. Viršutinėje eilėje pateikti originalūs vaizdai, o apatinėje – atitinkamos išraiškos, sudarytos remiantis Gramo matricos interpretacija [23] ir formulavimu. Nors šios vizualizacijos nėra generuotos siūlomo modelio, jos leidžia suprasti, kokio tipo informaciją fiksuoja Gramo matricos.

Apibendrindamos požymių atsakus per visas erdvines pozicijas, Gramo matricos išryškina tekstūrą, spalvų pasiskirstymą ir stilistinius dėsningumus, kartu susilpninant konkrečią erdvinę struktūrą. Todėl tokios reprezentacijos labiau atspindi žemo lygmens vizualinius požymius ir bendro pobūdžio stilistines detales, o ne tikslų objektų išdėstymą.

S.4 pav.: Iliustracinės Gramo matricos pagrindu sudarytos išraiškos. Viršutinė eilė - originalūs vaizdai; apatinė eilė - Gramo matricos pagrindu gautos išraiškos, išryškinančios tekstūrinę ir spalvinę informaciją. Originalūs vaizdai gauti iš EmoSet-118K duomenų rinkinio.

Ši savybė yra svarbi vizualinėms emocijos atpažinti, nes emocinės išraiškos interpretacija dažnai priklauso nuo bendros išvaizdos, spalvinės kompozicijos ir tekstūros, o ne nuo tikslios objektų padėties. Šie pavyzdžiai padeda geriau suprasti Gramo matricos pagrindu sudarytų išraiškų integraciją į siūlomą architektūrą.

S.4.2 Vertinimo metrikos ir kriterijai

Siekiant įvertinti modelių veiksmingumą ir jų gebėjimą atpažinti vizualines emocijas, šiame darbe naudojamos šios metrikos: tikslumas, preciziškumas, jautrumas ir $F1$ rodiklis. Šios metrikos apskaičiuojamos taip:

$$\begin{aligned} \text{Tikslumas} &= \frac{TP+TN}{TP+TN+FP+FN} \\ \text{Preciziškumas} &= \frac{TP}{TP+FP} \\ \text{Jautrumas} &= \frac{TP}{TP+FN} \\ F1 &= \frac{2 \cdot \text{Preciziškumas} \cdot \text{Jautrumas}}{\text{Preciziškumas} + \text{Jautrumas}} \end{aligned} \tag{S.11}$$

Čia TP (teisingi teigiami atvejai) yra pavyzdžiai, teisingai priskirti

teigiamai klasei, o TN (teisingi neigiami atvejai) – pavyzdžiai, teisingai priskirti neigiamai klasei. FP (klaidingi teigiami atvejai) – tai neigiami pavyzdžiai, neteisingai priskirti teigiamai klasei, o FN (klaidingi neigiami atvejai) – teigiami pavyzdžiai, neteisingai klasifikuoti kaip neigiami.

Tikslumas nusako bendrą teisingų prognozių dalį. Preciziškumas įvertina, kokia dalis teigiamai klasifikuotų pavyzdžių iš tiesų yra teigiami, o jautrumas parodo, kokia dalis tikrųjų teigiamų pavyzdžių buvo teisingai atpažinta. $F1$ rodiklis subalansuoja preciziškumą ir jautrumą ir yra naudingas esant nesubalansuotam klasių pasiskirstymui.

S.4.3 Kontrastinių centrų nuostolio funkcija

Kontrastinių centrų nuostolio funkcija naudojama siekiant pagerinti išraiškos požymių struktūrą požymių erdvėje: tos pačios klasės pavyzdžių požymių vektoriai artinami prie savo klasės centro, o skirtingų klasių centrai tolinami vieni nuo kitų.

Bendroji kontrastinių centrų nuostolio funkcija apibrėžiama taip:

$$L_{\text{contr}} = L_{\text{intra}} + \lambda L_{\text{inter}}, \tag{S.12}$$

kur L_{contr} yra bendra kontrastinių centrų nuostolio funkcija, L_{intra} yra vidinės klasės sutraukimo (angl. intra-class compactness) komponentas (sudėtinis dėmuo), L_{inter} yra tarpusavio klasių atskyrimo komponentas (angl. inter-class separability), o λ yra svorio hyperparametras, kontroliuojantis tarpusavio klasių atskyrimo įtaką.

Vidinės klasės sutraukimo komponentas apibrėžiamas taip:

$$L_{\text{intra}} = \frac{1}{N} \sum_{k=1}^{N} \left\| \mathbf{x}_k - \mathbf{c}_{y_k} \right\|^2, \tag{S.13}$$

kur N yra imčių skaičius, $\mathbf{x}_k \in \mathbb{R}^d$ yra k-ojo elemento požymių vektorius, $y_k \in \{1, \ldots, C\}$ yra k-ojo pavyzdžio klasės pažymėjimas, $\mathbf{c}_{y_k} \in \mathbb{R}^d$ yra klasei y_k priklausantis centras, C yra klasių skaičius, d yra požymių vektoriaus dimensija, o $\|\cdot\|$ žymi Euklidinę normą.

Šis komponentas mažina atstumą tarp požymių vektoriaus ir jį atitinkančio klasės centro.

Tarpusavio klasių atskyrimo komponentas apibrėžiamas taip:

$$L_{\text{inter}} = \frac{1}{C(C-1)} \sum_{i=1}^{C} \sum_{j \neq i} \max(0,\, m - \|\mathbf{c}_i - \mathbf{c}_j\|)^2, \tag{S.14}$$

kur $\mathbf{c}_i, \mathbf{c}_j \in \mathbb{R}^d$ yra atitinkamai i-osios ir j-osios klasės centrai, $m > 0$ yra klasės centrų atstumo apribojimo hyperparametras, nusakantis norimą mažiausią atstumą tarp skirtingų klasių centrų, o $\max(0, \cdot)$ reiškia, kad baudavimas (angl. penalty) taikomas tik tada, kai atstumas tarp atitinkamų centrų yra mažesnis už m.

Šis komponentas didina atstumus tarp skirtingų klasių centrų ir taip gerina klasių atskiriamumą požymių (reprezentacinėje) erdvėje.

Klasės centrai $\mathbf{c}_1, \ldots, \mathbf{c}_C$ yra apmokomi parametrai, inicializuojami atsitiktinai ir mokymo metu atnaujinami kartu su kitais modelio parametrais.

Bendra nuostolio funkcija tinklui apmokyti apibrėžiama taip:

$$L_{\text{total}} = L_{\text{entr}} + \beta L_{\text{contr}}, \tag{S.15}$$

kur L_{total} yra bendra nuostolio funkcija, L_{entr} yra praretintos kategorinės kryžminės entropijos nuostolio funkcija, L_{contr} yra kontrastinių centrų nuostolio funkcija, o β yra svertinis koeficientas - hiperparametras, kontroliuojantis kontrastinių centrų nuostolio funkcijos įtaką bendram apmokymui.

S.4.4 Priešingų sentimentų rodiklis

Be standartinių klasifikavimo metrikų, mes siūlome papildomą – priešingų sentimentų top-2 rodiklį (angl. cross-sentiment top-2 rate). Šis rodiklis yra skirtas modelio pirmų dviejų aukščiausių tikimybinių prognozių vidiniam nuoseklumui įvertinti. Skirtingai nuo tikslumu paremtų metrikų, kurios lygina prognozes su tikraisiais žymėjimais, šis rodiklis vertina,

ar modelis aukštą pasitikėjimą priskiria emocijų klasėms, priklausančioms priešingoms sentimentų grupėms (teigiama ir neigiama). Maža priešingų sentimentų rodiklio reikšmė rodo patikimesnį sentimentų elgesį, o didelė reikšmė gali atspindėti neapibrėžtumą ar nenuoseklumą išmoktoje išraiškoje.

EmoSet-118K ir FI-8 duomenų rinkiniai grindžiami Mikels emocijų ratu [49], todėl galima nustatyti, ar dvi aukščiausios prognozės priklauso tai pačiai, ar skirtingoms sentimentų grupėms.

Formaliai šis rodiklis apibrėžiamas taip. Kiekvienam įvesties pavyzdžiui (vaizdui) $\mathbf{x}_i$ tegu (c_1, p_1) ir (c_2, p_2) žymi pirmos ir antros pagal dydį prognozuotų emocijų klasių indeksus bei jų tikimybes. Apibrėžiame žymėjimą

$$f : \{0, 1, \ldots, 7\} \;\to\; \{\text{teigiama}, \text{neigiama}\},$$

kur klasės $\{0, 1, 2, 3\}$ priskiriamos teigiamai, o $\{4, 5, 6, 7\}$ – neigiamai sentimentų grupei.

Pavyzdys laikomas priešingu pagal sentimentą top-2 prognozėse, jei $f(c_1) \neq f(c_2)$. Tegu $\theta \in [0, 1]$ žymi pasitikėjimo (angl. confidence) slenkstį. Vertinami tik tie pavyzdžiai, kurių abi didžiausios tikimybės tenkina sąlygą

$$p_1 \geq \theta \quad \text{ir} \quad p_2 \geq \theta. \tag{S.16}$$

Jei N_θ žymi tokių tinkamų pavyzdžių skaičių, priešingų sentimentų top-2 rodiklis apibrėžiamas taip:

$$\text{Priešingas}(\theta) \;=\; \frac{1}{N_\theta} \sum_{i=1}^{N_\theta} \mathbf{1}[\, f(c_{1i}) \neq f(c_{2i}) \,], \tag{S.17}$$

čia $\mathbf{1}[\cdot]$ yra indikatoriaus funkcija.

Galimos kelios šio rodiklio interpretacijos:

- jei priešingų sentimentų rodiklis yra didelis, modelis gali būti neapibrėžtas dėl emocijų poliariškumo, o tai gali rodyti mažesnį patikimumą;

- jei rodiklis mažas, modelis linkęs nuosekliai atskirti sentimentus, kas rodo vidinį išraiškos nuoseklumą.

Gali kilti klausimas, kodėl verta taikyti prognozės tikimybės slenkstį. Slenkstis naudojamas ne tam, kad būtų fiksuojamos žemos tikimybės prognozės, bet tam, kad būtų išskirti atvejai, kai modelis top-2 klasėms vienu metu priskiria dideles tikimybes. Tokie atvejai rodo realų dviprasmiškumą, kai modelis nėra tikras, kuriai iš priešingų sentimentų grupių priskirti vaizdą. Taikant slenkstį top-2 tikimybėms, vertinamos tik reikšmingos dviprasmiškos situacijos, o atsitiktinės žemos tikimybės prognozės atmetamos.

S.4.5 Naudoti duomenų rinkiniai

- WEBEmo duomenų rinkinys sudarytas iš bendro pobūdžio emocijų vaizdų, surinktų iš viešai prieinamų internetinių šaltinių [61]. Šiame darbe iš jo sukonstruotas dviejų klasių duomenų rinkinys, kuriame atskiriami liūdesio vaizdai ir kitoms emocijoms priskirti vaizdai. Po vaizdų su tekstu pašalinimo ir papildomo imčių subalansavimo gautas galutinis WEBEmo liūdesio rinkinys, sudarytas iš 61074 vaizdų.

- FI-8 duomenų rinkinys sudarytas iš Flickr ir Instagram platformose surinktų vaizdų, naudojant aštuonias Mikels emocijų kategorijas [91]. Kiekvieną vaizdą vertino penki vertintojai, o į galutinį rinkinį buvo įtraukti tik tie vaizdai, kurių žyma surinko daugiau nei tris sutampančius balsus iš penkių. Galutinis FI-8 dydis yra 23308 bendro pobūdžio emocijų vaizdai.

- EmoSet-118K yra rankiniu būdu sužymėtas didesnio EmoSet rinkinio poaibis, apimantis 118102 vaizdus [89]. Vaizdai suskirstyti į aštuonias Mikels modelio emocijų kategorijas, o kiekvieną vaizdą vertino dešimt apmokytų vertintojų. Tikrasis atskiro vaizdo

žymėjimas buvo priskiriamas tik tada, kai dėl jo sutikdavo daugiau kaip septyni iš dešimties vertintojų.

- CAER-S yra iš didesnio CAER rinkinio sudarytas vaizdų rinkinys, sukonstruotas iš 79 televizijos serialų surinktos medžiagos [40]. Iš šio rinkinio atrinkti pavieniai vaizdai ir suformuotas septynių emocijų kategorijų rinkinys. Bendras CAER-S dydis yra 69999 vaizdai.

S.5 EKSPERIMENTAI IR REZULTATAI

Šiame skyriuje pateikiamas siūlomos vizualinių emocijų atpažinimo sistemos įvertinimas. Pirmiausia kiekybiškai įvertinamas pagalbinių Gramo matricos modulių indelis, lyginant bazinę architektūrą su konfigūracijomis, turinčiomis skirtingą lygiagrečiai prijungtų Gramo matricos modulių skaičių, naudojant standartines klasifikavimo metrikas ir įvairius vizualinių emocijų duomenų rinkinius.

Toliau vertinamas kontrastinių centrų nuostolio funkcijos integravimo poveikis, analizuojant tiek prognozavimo kokybę, tiek išraiškų struktūrą, taikant klasterių kokybės vertinimo metrikas (ARI, NMI, ASR) bei UMAP vizualizacijas. Galiausiai siūlomo modelio praktinis taikymo atvejis demonstruojamas atvejų analizėse su meno kūriniais ir WikiArt vaizdais, įskaitant nuoseklumo vertinimą, pagrįstą top-2 priešingų sentimentų matu.

Eksperimentai ir jų rezultatai publikuoti recenzuojamuose straipsniuose [A1, A2, B1].

S.5.1 Gramo matricos modulių indelio vertinimas

Eksperimentinio tyrimo tikslas – palyginti siūlomą modelį su baziniu modelio atveju, naudojant anksčiau apibrėžtas metrikas ir sužymėtų emocijas perteikiančių vaizdų duomenų rinkinius. Taip pat siekiama nustatyti tinkamą Gramo matricos modulių skaičių.

S.1 lentelė: Tikslumas WEBEmo liūdesio testavimo rinkinyje (vidurkis per 3 bandymus) ir palyginimas su bazine architektūra. Bazinis modelis atitinka EfficientNetV2S tinklą.

Tinklas	Tikslumas (%)	SD
Bazinis modelis	81,308	±0,548
Zhang *et al*. [96]	81,313	±0,186
2 Gramo matricos moduliai	82,313	±0,239
3 Gramo matricos moduliai	82,171	**±0,128**
4 Gramo matricos moduliai	**82,520**	±0,224

S.1 lentelėje pateiktas vidutinis tikslumas ir standartinis nuokrypis (SD). Siūlomas modelis pralenkia bazinį modelio atvejį maždaug 1,2 % aukštesniu tikslumu. Naudojant du Gramo matricos modulius ($v = 2$), gaunamas šiek tiek geresnis rezultatas nei naudojant tris modulius ($v = 3$). Keturių Gramo matricos modulių taikymas taip pat duoda gerų rezultatų.

Taip pat įvertintas Zhang et al. [96] modelis, naudojant jo straipsnyje aprašytą mokymo konfigūraciją (60 epochų). Mūsų bazinis modelis ir Gramo matricos modulių variantai buvo mokomi 20 epochų, nes tokio mokymo trukmės pakako konvergavimui pasiekti. Esant šioms sąlygoms, EfficientNetV2S pasiekė didesnį tikslumą nei Zhang et al. [96] modelis, kurio karkasas yra ResNet50.

Pažymėtina, kad mažiausias standartinis nuokrypis gautas, kai $v = 3$. Tai leidžia daryti prielaidą, kad didesnis Gramo modulių skaičius gali prisidėti prie rezultatų stabilumo ir bendro tinklo veikimo patikimumo. Tačiau $v = 4$ atveju standartinis nuokrypis taip pat mažesnis nei bazinio modelio, o pasiektas tikslumas – tik sąlyginai didesnis.

S.2 lentelėje pateikiami klasifikavimo rezultatai, apskaičiuoti kaip trijų bandymų vidurkiai. Palyginti su baziniu modeliu, visi Gramo matricos modulių konfigūracijos variantai pagerina kitų klasės tikslumą, jautrumą ir $F1$ rodiklį. Didžiausias kitų klasės $F1$ rodiklis pasiektas su $v = 1$ moduliu ($0,8405 \pm 0,0024$), o didžiausias tikslumas – su $v = 4$ moduliais ($0,8353 \pm 0,0059$).

Liūdesio klasėje visos Gramo modulių konfigūracijos pagerina tikslumą ir $F1$ rodiklį, palyginti su baziniu modeliu, o didžiausias jautrumas

S.2 lentelė: Tikslumo (precision), jautrumo (recall) ir $F1$ rodiklio rezultatai apskaičiuoti kaip 3 bandymų vidurkiai su standartiniais nuokrypiais. Bazinis modelis – EfficientNetV2S tinklas.

Modelis	**Kitos klasės**			**Liūdesys**		
	Tikslumas	Jautrumas	$F1$	Tikslumas	Jautrumas	$F1$
Bazinis modelis	0,8219 ± 0,0069	0,8359 ± 0,0053	0,8288 ± 0,0047	0,8024 ± 0,0056	0,7862 ± 0,0099	0,7942 ± 0,0066
1 modulis	0,8336 ± 0,0018	**0,8476 ± 0,0037**	**0,8405 ± 0,0024**	**0,8165 ± 0,0038**	0,8004 ± 0,0023	**0,8084 ± 0,0025**
2 moduliai	0,8313 ± 0,0036	0,8446 ± 0,0036	0,8379 ± 0,0021	0,8132 ± 0,0032	0,7978 ± 0,0056	0,8054 ± 0,0030
3 moduliai	0,8312 ± 0,0021	0,8415 ± 0,00053	0,8363 ± 0,00094	0,8102 ± 0,00046	0,7984 ± 0,0032	0,8042 ± 0,0017
4 moduliai	**0,8353 ± 0,0059**	0,8434 ± 0,0057	0,8393 ± 0,0017	0,8131 ± 0,0041	**0,8037 ± 0,0095**	**0,8084 ± 0,0036**

S.3 lentelė: Rezultatai priklausomai nuo β; WEBEmo liūdesys testavimo rinkinys.

β	Tikslumas ↑	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
1,2	0,8284	0,1402	0,2198	**0,2801**

pasiektas su $v = 4$ moduliais ($0,8037 \pm 0,0095$). Apskritai stebimas pagerėjimas abiejose klasėse, tačiau geriausia konfigūracija priklauso nuo pasirinktos metrikos: $v = 1$ maksimalizuoja kitų klasės $F1$ rodiklį, o $v = 4$ pasiekia didžiausią liūdesio klasės jautrumą.

S.5.2 Metrikų palyginimas

Kito eksperimento tikslas – ištirti emocijų išraiškas, išmoktas tinklo. Naudojant daugiamačių požymių vektorius siekiama įvertinti kontrastinių centrų nuostolio funkcijos integravimo veiksmingumą taikant anksčiau įvardytas metrikas.

S.4 lentelė: Rezultatai priklausomai nuo β; FI-8 testavimo rinkinys.

β	Tikslumas ↑	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
1,2	0,7138	0,4682	**0,4755**	**0,0728**

S.3 lentelėje pateikti modelio veikimo rezultatai WEBEmo liūdesio testavimo rinkinyje. Čia hyperparametras β kontroliuoja kontrastinių centrų nuostolio baudavimo (angl. penalty) stiprumą. Atvejis $\beta = 0$ atitinka bazinį modelį, kai naudojama tik kategorinė kryžminės entropijos nuostolio funkcija (žr. 3.1). Didžiausias tikslumas yra pasiektas esant $\beta = 1,0$, t. y. apie 0,7 % daugiau nei baziniame modelyje. ARI, NMI ir ASR rodikliai rodo, kad integravus kontrastinių centrų nuostolio funkciją pagerėja priešpaskutinio sluoksnio gaunamų klasterių kokybė.

S.4 lentelėje pateikiami FI-8 testavimo rinkinio rezultatai. Didžiausias tikslumas pasiektas esant $\beta = 0,5$ (apie 1,8 % pagerėjimas). Aukščiausi ARI ir NMI bei mažiausias ASR gauti esant $\beta = 1,2$. Tai rodo, kad kontrastinių centrų nuostolio integravimas gerina klasterių atskyrimą ir išraišką.

S.5 lentelėje pateikti EmoSet-118K testavimo rinkinio rezultatai. Didžiausias tikslumas ir geriausi ARI bei NMI rodikliai pasiekti esant $\beta = 1,0$. Mažiausias ASR matomas esant $\beta = 1,2$. Tokie rezultatai patvirtina, kad kontrastinių centrų nuostolio integravimas pagerina požymių erdvinę struktūrą.

Apibendrinant galima teigti, kad kontrastinių centrų nuostolio funkcijos integravimas nuosekliai gerina išmoktų požymių išraiškų struktūrą ir klasių atskiriamumą. Klasifikavimo tikslumo pagerėjimas kai kuriais atvejais yra nedidelis, todėl pagrindinis šios nuostolio funkcijos poveikis pasireiškia geresne požymių erdvės struktūra.

S.5 lentelė: Rezultatai priklausomai nuo β; EmoSet-118K testavimo rinkinys.

β	Tikslumas ↑	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
1,2	**0,7886**	0,5093	0,5906	**0,0759**

Kitas hyperparametras, skirtas emocijų klasių atskiriamumo nuostolio komponento L_{inter} svoriui reguliuoti, yra λ (žr. 4.21 lygtį). Didžiausias tikslumas pasiektas naudojant šias λ reikšmes: FI-8 atveju $\lambda = 100$, EmoSet-118K atveju $\lambda = 100$, o WEBEmo atveju $\lambda = 5$. Pastebima, kad optimali λ reikšmė skirtinguose duomenų rinkiniuose skiriasi. Vis dėlto, didėjant λ reikšmėms, jos įtaka rezultatams nėra labai ryški: praktikoje pakanka parinkti λ maždaug apie 5.

S.6 lentelė: Agreguotas palyginimas su baziniais modeliais trijuose testavimo rinkiniuose.

WEBEmo liūdesio	Tikslumas, %	ARI ↑	NMI ↑	ASR ↓
Bazinis modelis	0,8214	0,0987	0,1884	0,8797
Siūlomas modelis	**0,8374**	**0,1689**	**0,2293**	**0,0665**
EmoSet-118K	Tikslumas	ARI	NMI	ASR
Bazinis modelis	0,7794	0,3320	0,4550	0,3567
Siūlomas modelis	**0,8046**	**0,5988**	**0,6241**	**0**
FI-8	Tikslumas	ARI	NMI	ASR
Bazinis modelis	0,6968	0,3907	0,4337	0,8262
Siūlomas modelis	**0,7188**	**0,4592**	**0,4628**	**0,0646**

S.6 lentelėje pateikiamas geriausių konfigūracijų apibendrinimas kiekvienam duomenų rinkiniui, parinkus kontrastinių centrų nuostolio funkcijos hyperparametrus, įskaitant svertinį koeficientą β, ribą m ir λ.

Kiekvienam duomenų rinkiniui pateikti rezultatai atitinka tinkamiausią hyperparametrų parinkimą, nustatytą pagal ankstesniuose šio skyriaus eksperimentuose gautus rezultatus.

Palyginti su baziniais modeliais, kurie apmokyti nenaudojant kontrastinių centrų nuostolio funkcijos, siūlomas metodas nuosekliai pagerina klasifikavimo tikslumą ir išmoktų požymių išraiškų kokybę, ką rodo didesni ARI ir NMI rodikliai bei mažesnės ASR reikšmės. Šie rezultatai rodo, kad parinktos hyperparametrų konfigūracijos leidžia modeliui išmokti kompaktiškesnius ir geriau atskirtus emocijų klasterius požymių išraiškos erdvėje. Apibendrinti rezultatai patvirtina, kad kontrastinių centrų nuostolio integravimas, parinkus tinkamus hyperparametrus, lemia stabilesnes emocijų išraiškas ir geresnį klasių atskyrimą keliuose vizualinių emocijų duomenų rinkiniuose.

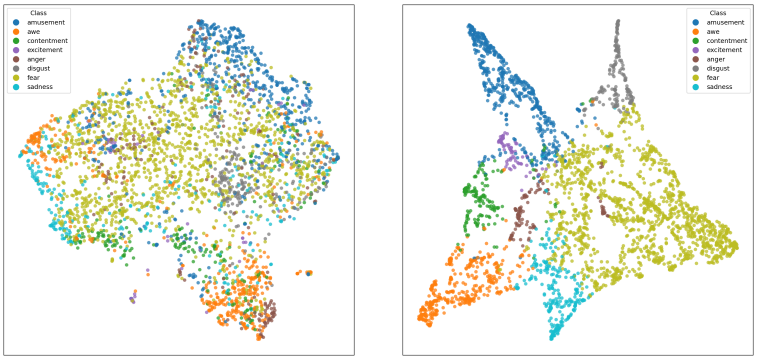
S.5.3 WikiArt emocijos

Vienas iš galimų sukurto vizualinių emocijų prognozavimo modelio praktinių taikymo atvejų yra meno kūrinių vertinimas. Literatūroje trūksta tyrimų, kuriuose būtų išsamiai aprašytas emocijų atpažinimo modelių praktinis taikymas. Todėl vizualinių emocijų atpažinimo taikymo analizė gali suteikti vertingų įžvalgų.

Kitas vizualinių emocijų duomenų rinkinys – WikiArt emocijos [50]. Šį rinkinį sudaro 4105 meniniai vaizdai, sužymėti pagal emocijas, kurias jie sukelia stebėtojui. Analizei atrinkti 3176 nesidubliuojantys vaizdai, priklausantys trims sentimentų grupėms: neigiamai, teigiamai ir neutraliai (mišriai).

S.5 paveiksle pateikiamas bazinio ir siūlomo modelio vizualizacijos rezultatų palyginimas. Čia UMAP metodu vizualizuojami 3176 požymių vektoriai. (b) atveju taškų pasiskirstymas yra aiškesnis ir geriau išlaikoma struktūra nei bazinio modelio atveju (a). Tai rodo, kad patobulinimai išlieka ir kituose vizualinių emocijų duomenų rinkiniuose.

Kaip aprašyta ankstesniame skyriuje, svarbu įvertinti modelio vidinį nuoseklumą. Tam pasitelkiamas modelio priešingų sentimentų top-2



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

S.5 pav.: Prognozių vizualizacija nesužymėtame WikiArt emocijų duomenų rinkinyje.

prognozių rodiklis (apibrėžtas lygtimi S.17). Šio eksperimento tikslas – patikrinti, ar patobulinta išraiškos požymių struktūra sumažina sentimentų lygmens painiavą. Šiai analizei nereikalingi tikrosios klasės žymėjimai, todėl galima įvertinti išraiškos vidinį užtikrintumą.

S.7 lentelė: Top-2 priešingų sentimentų įvertinimas WikiArt emocijų rinkinyje (priešingų, %).

Be slenksčio	Tinkamų	teig.-neig.	neig.-teig.	teig.-teig.	neig.-neig.	Priešingų
Bazinis	3176	652	664	518	1342	41,4
Pasiūlytas	3176	290	327	891	1668	**19,4**
Su slenksčiu $\theta = 0{,}3$	Tinkamų	teig.-neig.	neig.-teig.	teig.-teig.	neig.-neig.	Priešingų
Bazinis	552	122	121	85	231	43,5
Pasiūlytas	277	22	33	59	163	**19,9**

S.7 lentelėje pateikti top-2 priešingų sentimentų prognozių rezultatai. Lyginamas bazinis ir siūlomas modeliai. Tinkamų stulpelis nurodo vaizdų skaičių, kuriems tiek top-1, tiek top-2 tikimybės yra $\geq \theta$ (jei taikomas slenkstis). Siūlomas modelis pasiekia daug mažesnį priešingų sentimentų rodiklį už bazinį modelį. Mažesni priešingų sentimentų

įverčiai rodo geresnį emocinių išraiškų atskiriamumą. Taikant slenkstį $\theta = 0.3$, analizuojami tik tie atvejai, kai modelis pakankamai užtikrintai vertina abi aukščiausias prognozes. Taip atmetami atvejai, kai antrosios klasės tikimybė yra maža, ir išryškinamos situacijos, kai modelis iš tiesų dvejoja tarp dviejų alternatyvų.

S.8 lentelė: Top-2 priešingų sentimentų prognozių įvertinimas EmoSet-118K testavimo rinkinyje (priešingų, %).

Be slenksčio	Tinkamų	teig.-neig.	neig.-teig.	teig.-teig.	neig.-neig.	Priešingų
Bazinis	17716	1710	1710	8983	4726	22,6
Pasiūlytas	17716	507	738	10227	6244	**7,0**
Su slenksčiu $\theta = 0{,}3$	Tinkamų	teig.-neig.	neig.-teig.	teig.-teig.	neig.-neig.	Priešingų
Bazinis	2330	191	179	1442	518	15,9
Pasiūlytas	1420	73	59	1022	266	**9,3**

S.8 lentelėje pateikta top-2 priešingų sentimentų prognozių EmoSet-118K testavimo rinkiniuose. Modelis apmokytas atitinkamai to paties rinkinio apmokymo aibėje. Mūsų siūlomas modelis pastebimai pasiekia mažesnį priešingų sentimentų rodiklį už bazinį modelį, o tai rodo sumažėjusią painiavą tarp teigiamų ir neigiamų sentimentų grupių. Šis pagerėjimas stebimas tiek be slenksčio, tiek taikant prognozių klasifikavimo slenkstį $\theta = 0,3$.

Apibendrinant galima teigti, kad modifikuota mokymo strategija, integruojanti kontrastinių centrų nuostolio funkciją, pagerina sentimentų lygmens atskiriamumą ir sustiprina išmoktų išraiškų vidinį nuoseklumą.

S.6 BENDROSIOS IŠVADOS

Disertacijoje sprendžiama vizualinių emocijų atpažinimo bendros paskirties vaizduose problema, kuriant ir vertinant CNN architektūrų bei mokymo metodų išplėtimus. Tyrimas orientuotas į emocinės išraiškos

atotrūkio mažinimą ir emocijų klasių atskiriamumo didinimą – esminius vizualinių emocijų analizės iššūkius.

Pagrindinės disertacijos išvados:

1. Emocijos gali būti išreikštos vaizdinių duomenų terpėje. Reikalingi automatizuoti metodai, galintys veiksmingai vertinti didelės apimties vizualinių emocijų duomenis. Mūsų siūlomas CNN pagrindu sukurtas modelis užtikrina patikimesnę emocijų klasifikaciją bendro pobūdžio vaizduose, gerinant išraiškos požymių struktūrą.

2. Eksperimentai parodė, kad Gramo matricos modulių integravimas į EfficientNetV2S bazinį tinklą gerina vizualinių emocijų klasifikavimo tikslumą visose tirtose konfigūracijose. Bazinis tinklas pasiekė $81{,}31\,\% \pm 0{,}548\,\%$, o siūlomas modelis su keturiais Gramo matricos požymių moduliais – $82{,}52\,\% \pm 0{,}224\,\%$, t. y. $1{,}2\,\%$ didesnį tikslumą WEBEmo liūdesio duomenų rinkinyje. Zhang et al. [96] pasiūlytas modelis pasiekė $81{,}313\,\% \pm 0{,}186\,\%$ tikslumą, todėl siūlomu metodu gauti dar geresni rezultatai, palyginti su ankstesniais Gramo matrica pagrįstais metodais. Be to, keturių Gramo matricos modulių konfigūracija pasižymėjo mažiausiu standartiniu nuokrypiu, o tai rodo ne tik didesnį tikslumą, bet ir stabilesnį modelio veikimą.

3. Kontrastinių centrų nuostolio funkcijos integravimas pagerino klasifikavimo tikslumą visuose vertintuose emocijų duomenų rinkiniuose. WEBEmo liūdesio duomenų rinkinyje tikslumas padidėjo nuo 82,14 % (bazinis modelis) iki 83,74 % (siūloma konfigūracija). FI-8 duomenų rinkinyje tikslumas padidėjo nuo 69,68 % iki 71,88 %, geriausių rezultatų pasiekta esant $\beta \approx 0{,}5$. EmoSet-118K duomenų rinkinyje tikslumas išaugo nuo 77,94 % iki 80,46 %, geriausių rezultatų gauta esant $\beta \approx 1$. Kontrastinių centrų optimizacijos įtraukimas davė reikšmingą pagerėjimą skirtinguose vizualinių emocijų duomenų rinkiniuose.

4. Klasterių analizė parodė, kad kontrastinių centrų nuostolis gerina išmoktos požymių erdvės struktūrą. Visuose vaizdinių emocijų duomenų rinkiniuose ARI ir NMI rodikliai didėja, o ASR mažėja –

tai rodo kompaktiškesnius emocijų klasterius ir mažesnį skaičių pavyzdžių, esančių arti kelių klasių centrų. EmoSet-118K duomenų rinkinyje ARI padidėjo nuo 0,332 iki 0,5988, NMI – nuo 0,455 iki 0,6241, o ASR sumažėjo iki 0. WEBEmo liūdesio duomenų rinkinyje ASR sumažėjo nuo 0,8797 iki 0,0665. Šie rezultatai atitinka dimensijų mažinimo ir išraiškų vizualizacijos išvadas. Siūlomas modelis rodo aiškesnį grupavimą ir ryškesnį emocijų poliariškumo atskyrimą, kai $\beta > 0$, o tai patvirtina, jog kontrastinių centrų optimizacija leidžia suformuoti aiškiau atskirtas emocijų išraiškas.

5. Siūlomas top-2 priešingų sentimentų rodiklis papildo modelio vertinimą, vertindamas, kaip dažnai dvi didžiausios modelio prognozės patenka į priešingas sentimentų grupes. WikiArt emocijų duomenų rinkinyje siūlomas modelis pasiekė gerokai mažesnį priešingų senti-mentų rodiklį už bazinį modelį tiek be prognozių slenksčio (nuo 41,4 % iki 19,4 %), tiek taikant slenkstį $\theta = 0{,}3$ (nuo 43,5 % iki 19,9 %). Tai rodo stabilesnį modelio užtikrintumą sentimentų lygmeniu ir patvirtina kontrastinių centrų optimizacijos veiksmingumą mažinant dviprasmiškumą tarp priešingų emocijų grupių.

Disertacijoje pasiektas tikslas – pasiūlyti ir empiriškai pagrįsti efektyvius CNN pagrįstus metodus vizualinėms emocijoms atpažinti. Ateities tyrimuose tikslinga plėtoti metodiką multimodalinės emocijų analizės kryptimi ir tirti praktinio pritaikomumo galimybes sveikatos priežiūros srityje.

CURRICULUM VITAE

Modestas Motiejauskas received his BS in Information Technologies from Vilnius University in 2018 and his MS in Computer Modelling from Vilnius University in 2021. His current research interests include deep neural networks, computer vision, and visual image emotion analysis.

Notes

Modestas Motiejauskas
Convolutional Neural Network Architectures and Training Improvements:
Visual Emotion Recognition Case
Doctoral Dissertation
Natural Sciences
Informatics (N 009)
Thesis Editor: Zuzana Šiušaitė

Konvoliucinių neuroninių tinklų architektūrų ir mokymo gerinimas: emocijų atpažinimo vaizduose atvejis
Daktaro disertacija
Gamtos mokslai
Informatika (N 009)
Santraukos redaktorė: Jorūnė Rimeisytė-Nekrašienė

Vilniaus universiteto leidykla
Saulėtekio al. 9, III rūmai, LT-10222 Vilnius
El. p. info@leidykla.vu.lt, www.leidykla.vu.lt
bookshop.vu.lt, journals.vu.lt
Tiražas 20 egz.