A thesis for the degree of Doctor of Philosophy

A Study on Machine Learning for Personalized Prediction of Human Perception toward Visual Stimuli

視覚刺激に対する人間の知覚を
個別予測するための機械学習に関する研究

Yuya Moroto

Graduate School of Information Science and Technology

Hokkaido University

March, 2024

Abstract

This thesis summarizes studies on the construction of machine learning models specific to personalized prediction of human perception toward visual stimuli.

Machine learning has attracted significant attention in assisting humans due to its high potential and continues to respond to expectations in various fields. Specifically, after the development of deep learning technologies such as convolutional neural networks and recurrent neural networks, machine learning models can solve more complex tasks by learning a large amount of data. Recent studies on machine learning have progressed to the foundation models such as contrastive language-image pre-training and generative pre-trained transformer, and researchers have focused on the way to construct models that can effectively learn big data and conduct several tasks in a single model. Namely, they aim to develop the generalized model. Although this direction may be one advancement of machine learning, another important direction is the development of machine learning models that can be tuned for each individual from the perspective of human assistance. For instance, user satisfaction in video-sharing services can be improved by personalizing the multimedia content recommender system. Therefore, the personalization of machine learning can be an effective direction of advancement.

The person-specific information is needed as a clue for training machine learning models to suit each individual. One of the person-specific information is the biological information obtained from humans. Here, to introduce such information into the machine learning models, human perception should be mediated as in the actual human information processing. However, it is difficult to directly implement them to existing models in various tasks such as content recommendation and information retrieval since machine learning just recognizes the patterns in the inputs and outputs and may ignore human perception. Hence, studies on predicting human perception have been conducted to indirectly personalize machine learning. Concretely, previous studies have predicted emotion and attention as human perception from brain activity and gaze data as the data representing biological information (hereafter, biological data). In these studies, although machine learning models have been used as prediction models, these models do not necessarily consider the properties specific to biological data since their architectures are designed not specifically for biological data. In contrast to the general data in the fields of computer vision and natural language processing, biological data are difficult to handle due to their unique properties such as individual differences. Therefore, there are great demands for rethinking the machine learning models suitable for biological data.

This thesis focuses on three perspectives related to the inherent properties of biological data. The first perspective is the data volume obtained from each individual. Biological data varies widely among individuals, and data obtained from various individuals are difficult to handle in a uniform manner. Hence, the machine learning models need to be trained from the limited amount of data for reflecting on individual differences. The next perspective is the relationship

between stimuli and their human response. Humans constantly receive a variety of stimuli and perceive them in their daily lives, and biological data reflect on such stimuli. To effectively predict human perception, not only biological data but also the contents of stimuli should be considered. Finally, the third perspective is mutual complementation through the collaborative use of several types of biological data. Advancements in sensor technologies enable the easy and simultaneous acquisition of various types of biological data. Each type of biological data represents a different aspect of the human response, and the human perception can be more precisely predicted by collaboratively using them than one of them alone.

The purpose of this thesis is to construct machine learning models that can predict personalized human perception by incorporating the above perspectives. This thesis targets the human perception toward visual stimuli since several studies show that visual information is the most important to humans. Concretely, this thesis mainly tackles three themes to construct the machine learning models incorporating the above perspectives, respectively. First, to address the problem of the data volume, we focus on the similarities of biological data between individuals. In the case of predicting human attention toward visual stimuli such as images, we propose a new method for detecting the individuals with biological data patterns similar to those of the target individual. Moreover, we construct the machine learning model using the data obtained from similar individuals for predicting the perception of the target individual. Secondly, for analyzing the relationship between visual stimuli and biological data, we focus on the construction of the uniform representation of visual contents and gaze data including the region watched by the individual. Finally, we newly propose the feature integration methods for treating several types of biological information since biological data are pre-processed for calculating features suitable for each type of data before inputting machine learning models, generally. Then, when calculating the features of gaze data, we adopt the representation based on the second perspective for considering both visual contents and biological data. In this way, we newly proposed machine learning models suitable for biological data and indicate the effectiveness of focusing on the above inherent perspective.

This thesis consists of six chapters. Chapter 1 describes the research background and the proposition of this thesis. Chapter 2 describes the related works and their problems to be solved in this thesis. Chapter 3 presents methods for few-shot personalized saliency prediction, which is the task predicting regions in images gazed at by individuals. Chapters 4 and 5 focus on human emotions as perceptions. Chapter 4 presents the methods for classifying images into emotional categories using gaze data. Chapter 5 presents the methods for multi-modal human emotion recognition based on various types of biological information. Finally, Chapter 6 concludes this thesis and clarifies the future directions.

In summary, this thesis presents several machine learning methods for personalized prediction of human perception toward visual stimuli. For constructing the machine learning models specific to personalized prediction of human perception, the proposed methods incorporate the similarities of biological data between individuals and mutual complementation between different types of biological information. Furthermore, we confirm the effectiveness of the proposed methods through empirical experimentation on datasets derived from personally acquired raw data and openly available datasets.

Acknowledgements

This doctoral dissertation comprises research conducted over approximately six years during my enrollment at Hokkaido University and its Graduate School. Its completion would not have been possible without the support and guidance I received from others.

In the pursuit of this research, I extend my sincere appreciation to Prof. Miki Haseyama for her unwavering guidance and encouragement. Beyond providing an enriching research environment, she facilitated numerous valuable opportunities, including participation in domestic and international conferences and the submission of academic papers, for which I am deeply thankful.

In crafting this dissertation, profound gratitude is extended to Specially Appointed Prof. Kenji Araki, Specially Appointed Prof. Yuji Sakamoto, and Prof. Yoshinori Dobashi, Prof. Takahiro Ogawa for their fruitful comments on my thesis and role as committee members.

Further appreciation is directed towards Associate Prof. Sho Takahashi of the Graduate School of Engineering at Hokkaido University, Associate Prof. Ryosuke Harakawa of the Department of Electrical and Electronic Information Engineering at Nagaoka University of Technology, Assistant Prof. Naoki Saito of the Office of Institutional Research at Hokkaido University, Specially Appointed Assistant Prof. Ren Togo of the Media Dynamics Laboratory at the Graduate School of Information Science and Technology, Hokkaido University, and Specially Appointed Associate Prof. Keisuke Maeda of Data-Driven Interdisciplinary Research Emergence Department at Hokkaido University.

Specifically, I would like to appreciate Prof. Ogawa for his guidance on the formulation and organization of the research, laying the foundation for my study. Moreover, I am deeply grateful for his generous advice, not only on research activities but also on personal matters, including discussions about pursuing a doctoral program. Additionally, I also would like to thanks to Dr. Maeda for his valuable guidance, not only in research endeavors but also in providing advice on

student life, despite my relative inexperience.

I would like to express my thanks to all of my colleagues at the Laboratory of Media Dynamics at the Graduate School of Information Science and Technology, Hokkaido University for their care and help.

Furthermore, during the my PhD research tenure, I would like to gratefully acknowledge the financial support provided by the Nitori International Scholarship Foundation, the Japan Educational Exchanges and Services (JEES) and Mitsubishi Corporation, the Sky Oura ICT Scholarship Foundation, the Japan Student Services Organization, and the Japan Society for the Promotion of Science.

Finally, I express my deepest gratitude to my family for their understanding of my career choices and substantial support, both economically and emotionally, throughout my academic journey. This dissertation stands as a proof to the collective contributions of these esteemed individuals and institutions.

謝辞

本研究は，著者が北海道大学および北海道大学大学院に在学した期間，約6年間にわたって行ったものである．この6年の間に多くの方々に御支援を頂いた.

本研究を行うにあたり，研究の遂行に対して終始御指導，鞭撻を頂きました長谷山美紀教授に深く感謝いたします．また，充実した研究環境を用意して頂けただけでなく，国内外の学会への参加や学術論文誌への投稿等，多数の有益な機会を提供頂きましたことに対しても深くお礼申し上げます.

本論文をまとめるにあたり，御助言を賜り，副査をお引き受けいただいた北海道大学大学院情報科学研究院 言語メディア学研究室 荒木健治 特任教授，北海道大学大学院情報科学研究院 メディア創生学研究室 坂本雄児 特任教授, 北海道大学大学院情報科学研究院 情報メディア環境学研究室 土橋宜典 教授，ならびに北海道大学大学院情報科学研究院 メディアダイナミクス研究室 小川貴弘 教授に深く感謝いたします.

また，北海道大学大学院工学研究院 先端モビリティ工学研究室 高橋翔 准教授，長岡技術科学大学 電気電子情報工学専攻 画像・メディア工学研究室 原川良介 准教授，北海道大学 総合IR本部 斉藤 直輝 助教，北海道大学大学院情報科学研究院 メディアダイナミクス研究室 藤後廉 特任助教，北海道大学 創成研究機構 データ駆動型融合研究創発拠点 前田圭介 特任准教授におきましては，本研究を遂行するにあたり，公私にわたり多大なる御助力を頂けましたことを深く感謝いたします.

小川教授におきましては，研究の立案方法やまとめ方等，研究の基礎を教えて頂きました．また，研究活動のみならず，博士後期課程への進学に関する相談等，公私にわたり寛大な心で御助言を頂きましたこと，深く感謝いたします．さらに，前田特任准教授におきましては，未熟な私に対しても，研究活動はもちろんのこと，学生生活におけるご助言等を頂きましたこと,深く感謝いたします.

著者の研究室所属期間中，多くの御協力を頂きました，北海道大学大学院情報

科学院情報科学専攻メディアダイナミクス研究室の諸先輩，同輩，ならびに後輩の皆様に感謝致します．皆様と共に切磋琢磨し合うことで，約6年にわたる私の研究を無事に遂行できました．

また，著者の研究期間中に経済的な支援を賜りました，公益財団法人似鳥国際奨学財団，公益財団法人日本国際教育支援協会ならびに三菱商事株式会社，公益財団法人Sky大浦ICT奨学財団，独立行政法人日本学生支援機構，独立行政法人日本学術振興会に深く感謝いたします．

最後に，私の進路選択に理解を示してくださり，在学期間中に経済面および精神面で多大な支援をしていただいた家族に深い感謝の意を表し謝辞とさせて頂きます．

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter shows the background, the proposition, and the organization of this thesis.

1.1 Background

Machine learning has attracted significant attention in assisting humans due to its high potential and continues to respond to expectations in various fields [1–3]. Specifically, after the development of deep learning technologies such as convolutional neural networks (CNNs) [4] and recurrent neural networks [5], machine learning models can solve more complex tasks by learning a large amount of data. Recent studies on machine learning have progressed to the foundation models such as contrastive language-image pre-training [6] and generative pre-trained transformer [7], and researchers have focused on the way to construct models that can effectively learn big data and conduct several tasks in a single model [8–10]. Namely, they aim to develop the generalized model. Although this direction may be one advancement of machine learning, another important direction is the development of machine learning models that can be tuned for each individual from the perspective of human assistance. For instance, user satisfaction in video-sharing services can be improved by personalizing the multimedia content recommender system [11, 12]. Therefore, the personalization of machine learning can be an effective direction of advancement.

The person-specific information is needed as a clue for training machine learning models to suit each individual. One of the person-specific information is the biological information ob-

tained from humans [13]. Here, to introduce such information into the machine learning models, human perception should be mediated as in the actual human information processing. It should be noted that the human perception is defined as the broad interpretation of the stimuli such as attention and emotion in this thesis. However, it is difficult to directly implement them to existing models in various tasks such as content recommendation [11, 12] and information retrieval [14, 15] since machine learning just recognizes the patterns in the inputs and outputs and may ignore human perception. Hence, studies on predicting human perception have been conducted to indirectly personalize machine learning [16–19]. Concretely, previous studies have predicted emotion and attention as human perception from brain activity and gaze data as the data representing biological information (hereafter, biological data). In these studies, although machine learning models have been used as prediction models, these models do not necessarily consider the properties specific to biological data since their architectures are designed not specifically for biological data. In contrast to the general data in the fields of computer vision and natural language processing, biological data are difficult to handle due to their unique properties such as individual differences. Therefore, there are great demands for rethinking the machine learning models suitable for biological data.

1.2 Proposition in this Thesis

This thesis focuses on three perspectives related to the inherent properties of biological data. The first perspective is the data volume obtained from each individual. Biological data varies widely among individuals, and data obtained from various individuals are difficult to handle in a uniform manner. Hence, the machine learning models need to be trained from a limited amount of data for reflecting on individual differences. The next perspective is the relationship between stimuli and their human response. Humans constantly receive a variety of stimuli and perceive them in their daily lives, and biological data reflect on such stimuli. To effectively predict human perception, not only biological data but also the contents of stimuli should be considered. Finally, the third perspective is mutual complementation through the collaborative use of several types of biological data. Advancements in sensor technologies enable the easy and simultaneous acquisition of various types of biological data. Each type of biological data

represents a different aspect of the human response, and the human perception can be more precisely predicted by collaboratively using them than one of them alone.

The purpose of this thesis is to construct machine learning models that can predict personalized human perception by incorporating the above perspectives. This thesis targets the human perception toward visual stimuli since several studies show that visual information is the most important to humans. Concretely, this thesis mainly tackles three themes to construct the machine learning models incorporating the above perspectives, respectively. First, to address the problem of the data volume, we focus on the similarities of biological data between individuals. In the case of predicting human attention toward visual stimuli such as images, we propose a new method for detecting the individuals with biological data patterns similar to those of the target individual. Moreover, we construct the machine learning model using the data obtained from similar individuals for predicting the perception of the target individual. Secondly, for analyzing the relationship between visual stimuli and biological data, we focus on the construction of the uniform representation of visual contents and gaze data including the region watched by the individual. Finally, we newly propose the feature integration methods for treating several types of biological information since biological data are pre-processed for calculating features suitable for each type of data before inputting machine learning models, generally. Then, when calculating the features of gaze data, we adopt the representation based on the second perspective for considering both visual contents and biological data. In this way, we newly proposed machine learning models suitable for biological data and indicate the effectiveness of focusing on the above inherent perspective.

1.3 Organization of this thesis

This thesis contains six chapters. The first chapter is this chapter, and the rest of this thesis is organized as below.

Chapter 2 describes the related works of visual saliency prediction, emotional categorical classification, and multi-modal human emotion recognition, and the most representative ones are listed. Besides, the problems of these works to be solved are clarified.

Chapter 3 presents three methods for the prediction of the personalized saliency map (PSM)

with a limited amount of training data, and consists of three chapters for presenting each method. Concretely, Chapter 3.1 presents the few-shot PSM prediction method based on adaptive image selection considering object and visual attention. This method focuses on the similarities of visual attention between individuals for the prediction of personalized salient regions in images from a limited amount of training data. To calculate such similarities, the images that individuals commonly gazed at are needed. Hence, the adaptive image selection module considering object and visual attention is proposed and introduced into the PSM prediction model in a simple manner. Next, Chapter 3.2 presents the few-shot PSM prediction method using individual similarity based on collaborative multi-output Gaussian process regression. This method is an extended version of the method proposed in Chapter 3.1. In the method presented in Chapter 3.1, the PSM is predicted by simply using similarities invariant to images. Then, this method incorporates the machine learning model with the similarities and the target image for predicting the PSM with varying similarities for each image. In this method, the Gaussian process regression-based model is adopted for considering the gaze uncertainty and data volume. Finally, Chapter 3.3 presents the few-shot PSM prediction method with similarity of gaze tendency using object-based structural information. In this method, the remaining problem of the method presented in Chapter 3.2, which is the collapse of structural information of images, is solved. For preserving the structural information, this method focuses on the object-based similarities of gaze tendency. Experiments with the open dataset showed a progressive improvement in performance in each chapter.

Chapter 4 presents two methods for gaze-based emotional category classification of images and consists of two chapters for presenting each method. Concretely, Chapter 4.1 presents an estimation of emotion labels via tensor-based spatiotemporal visual attention. In this method, a novel way to construct a uniform representation including visual contents and gaze data is proposed for effectively analyzing the relationship between visual stimuli and biological data. The constructed representations are the fourth-order tensors, and the machine learning-based tensor analysis is applied to them for estimating emotion labels for images. By confirming the performance of emotion label estimation, such representation is indicated to contain both the visual contents and gaze data. Chapter 4.2 presents tensor-based emotional category classification via visual attention-based heterogeneous CNN feature fusion. This method focuses on the feature

extraction from the representation presented in Chapter 4.2. CNN features, which are outputs of an intermediate layer of the pre-trained CNN, are well-known for their high representation ability. However, they do not necessarily have the high discrimination ability for our target domain, and this method uses multiple CNN features calculated from multiple CNN models by using tensor analysis-based feature fusion. Experimental results showed a progressive improvement in performance in each chapter.

Chapter 5 presents three methods for multi-modal human emotion recognition using several types of biological information and consists of three chapters for presenting each method. Concretely, Chapter 5.1 presents the human-centric emotion estimation method based on correlation maximization considering changes with time in visual attention and brain activity. This method simply focuses on the correlation-based feature integration treating several types of biological information. By using the canonical correlation analysis, heterogeneous features are transformed into the common feature spaces with properties of multiple input features. Transformed features are input to the simple machine learning model for predicting human perception. Chapter 5.2 presents the human emotion recognition method using multi-modal biological data based on time-lag considered correlation maximization. In human emotion recognition of visual stimuli, humans gather information through their eyes, which is subsequently processed in the brain. Visual stimuli perceived by human eyes undergo transmission to the brain through neurotransmitters, resulting in a time delay between gaze data and brain activity data. Hence, this method integrates features with considering the time lag between multiple biological data. Finally, Chapter 5.3 presents the multi-view variational recurrent neural network for human emotion recognition using multi-modal biological data. This method focuses on the other characteristics of biological data. Specifically, this method realizes feature integration with considering the following three characteristics: 1) the relationship between explicit and implicit information such as brain activity and gaze, 2) temporal changes associated with emotions recalled by humans, and 3) the potential impact of noises. For simultaneously considering them, the multi-view variational recurrent neural network is newly derived. Experiments on datasets derived from personally acquired raw data showed a progressive improvement in performance in each chapter.

Chapter 6 concludes this thesis and describes the future direction.

The methods presented in each chapter correspond to the research achievements at the end of this thesis. Chapters 3.1, 3.2, and 3.3 introduce the methods proposed in [A-2], [B-9], and [B-12], respectively. Chapters 4.1 and 4.2 introduce the methods proposed in [A-1] and [B-4]. Finally, Chapters 5.1, 5.2, and 5.3 introduce the methods proposed in [A-3], [B-11], and [B-18]. It should be noted that figures and tables in this thesis are taken from or partially modified from the corresponding references.

Chapter 2

Related Works

2.1 Introduction

This chapter shows the research related to this thesis. For realizing the personalized prediction of human perception toward visual stimuli, this thesis mainly focuses on three themes, the personalized saliency prediction, emotional category classification, and multi-modal human emotion recognition. Therefore, this chapter presents several studies relevant to the three tasks mentioned above. Concretely, Section 2.2 presents several studies relevant to saliency prediction. Section 2.3 presents several studies relevant to emotional category classification of images. Section 2.4 presents several studies relevant to multi-modal human emotion recognition. Next, Section 2.5 clarifies the problems to be solved in this thesis. In the end, Section 2.6 concludes this chapter.

2.2 Saliency Prediction

While this thesis addresses personalized saliency prediction, the field of image processing traditionally focuses on universal saliency prediction. Actually, personalized saliency prediction has been rarely studied since it is difficult to acquire gaze data including the gazed location in images. Instead, the universal saliency prediction has been studied for clarifying the human visual system and implementing them into computers. Although the map predicted by the universal saliency prediction model is called a saliency map, generally, the map predicted by the

personalized saliency prediction model is called a personalized saliency map (PSM), and the traditional saliency map is called a universal saliency map (USM) in this thesis. Besides, it should be noted that a saliency map means the generic term of PSMs and USMs. Hereafter, the previous works related to USM prediction and PSM prediction are summarized in Section 2.2.1 and Section 2.2.2, respectively. Finally, Section 2.2.3 provides an overview of metrics for evaluating the quality of predicted maps since PSM and USM require specific evaluation metrics distinct from general machine learning tasks.

2.2.1 Previous Works Related to USM Prediction

This section explains research on USM prediction. Hereafter, mathematical research and deep learning-based research are described as USM prediction research. Moreover, the reference that compare the prediction accuracy of these two types of USM prediction methods is introduced.

Mathematical Research

Reference [20]

Reference [20] proposed the first mathematical model for USM prediction by utilizing a Gaussian pyramid to represent a physiological model for human visual attention. This model calculates feature maps related to luminance, hue, and orientation components for multiple images obtained by applying the Gaussian pyramid to the input image. USMs are calculated through feature integration, incorporating normalization processes inspired by the receptive fields in retinal ganglion cells [21].

Reference [22]

Reference [22] calculates hand-crafted feature vectors similar to reference [20]. The method embeds the dissimilarity between target regions and their neighbors into a Markov chain-based graph and calculates the USMs for images using a method similar to random walk.

Reference [23]

Reference [23] focuses on extracting the foreground region of an image by applying

thresholding to frequency components obtained through discrete cosine transform. USMs are then predicted by summing the obtained foreground regions in the hue component direction after Gaussian blurring. This study primarily addresses objects in the foreground of images.

Reference [24]

Reference [24] calculates local features by extracting color and texture features from segmented regions of the image. Moreover, global features are obtained by calculating color distribution from the entire image, and these global and local features are used to calculate saliency scores. The final USM prediction is achieved through integration of these local and global saliency scores.

Deep Learning-based Research

Reference [25]

Reference [25] proposed the first deep learning-based method for USM prediction. The study constructs multiple Convolutional Neural Networks (CNNs) [4] with hierarchical structures inspired by biology. Moreover, the output values from these CNNs are integrated using Support Vector Machine, resulting in the final USM prediction.

Reference [26]

Reference [26] employs a CNN with five layers for USM prediction. Besides, the method presented in reference [26] was improved in 2016 [27], achieving higher accuracy in USM prediction by utilizing both shallow and deep features obtained from CNNs.

Reference [28]

Reference [28] addresses the problem of predicting salient regions that are not strongly associated with semantic contents in images. The study combines the output values from two pre-trained CNNs of different sizes, solving this problem and achieving accurate USM prediction.

Reference [29]

Reference [29] utilizes a Neural Network based on Generative Adversarial Network (GAN) [30]

for USM prediction. This method constructs a generator and a discriminator, with the generator performing USM prediction and the discriminator distinguishing between the predicted USM and the Ground Truth.

Comparison and Summary of Previous Research

In reference [31], five mathematical model-based and five deep learning-based USM prediction methods are compared and evaluated through experiments. The results demonstrate that, across multiple evaluation metrics, deep learning-based methods outperform mathematical model-based methods in terms of prediction accuracy. When comparing the average accuracy of deep learning-based and mathematical model-based methods, deep learning-based methods consistently exhibit higher average accuracy across all evaluation metrics, with statistical significance confirmed. Therefore, it is suggested that deep learning-based USM prediction methods are superior, although they may exhibit reduced accuracy for distorted images.

2.2.2 Previous Works Related to PSM Prediction

Reference [32]

In this work, the first dataset designed for PSM prediction has been constructed as an open dataset. This dataset encompasses gaze data and images captured when 30 experimental participants gazed at 1600 images. Building upon this dataset, a PSM prediction based on a multi-task CNN [33] has been proposed. Despite the substantial amount of gaze data collected from multiple participants, it was deemed insufficient for training deep learning models. To address this limitation, Xu *et al.* introduced a multi-task CNN that is capable of simultaneously predicting PSMs for multiple participants, thereby compensating for data volume and achieving highly accurate PSM prediction.

Reference [34]

This study further advanced PSM prediction by incorporating individual background information. While the method proposed in [32] allows for accurate PSM prediction, the problem arose concerning the scale of the network. To tackle this problem, this study

introduces a strategy employing both the original image and individual background information as inputs. This approach employs a unified network to learn gaze data obtained from all participants, addressing issues related to network size and data volume.

2.2.3 Evaluation Metrics for Predicted Saliency Map

Research on predicting USMs and PSMs employs diverse evaluation metrics. In reference [35], Bylinskii *et al.* systematically categorizes and elucidates these evaluation metrics, outlining their roles as below.

Area under ROC Curve (AUC)

Given that saliency maps depict fixation locations on images, saleincy prediction can be conceptualized as a classification problem. AUC evaluates the accuracy of saliency prediction based on the area under the Receiver Operating Characteristic (ROC) curve, which is a standard for signal detection in the field of signal processing. This metric allows for precise evaluation focusing on locations on images.

Shuffled AUC (sAUC)

sAUC, an extension of AUC, was devised to address biases in evaluating saliency maps. Since humans exhibit a center bias and gazed at the central part of images [36], models that predict only the central region as the fixation location tend to receive inflated AUC scores. sAUC mitigates this bias by randomly sampling fixation locations from other images for eliminating center bias effects.

Normalized Scanpath Saliency (NSS)

NSS is an evaluation metric devised for assessing saliency prediction. While AUC and sAUC evaluate saliency maps based on the accurate prediction of high Ground Truth values, they tend to yield high evaluation scores even when large errors occur in other regions. However, a model that predicts high values for regions not actually gazed at may not be effective since the relative gaze intensity is important information. Therefore, NSS addresses this issue by normalizing the predicted saliency map.

Pearson's Correlation Coefficient (CC)

CC, a statistical measure widely used across various fields, evaluates the overall similarity between the predicted saliency map and the Ground Truth. In contrast to metrics that explicitly penalize false positives and false negatives, CC provides a uniform value to evaluate the overall similarity between the maps. While a low CC value makes it challenging to discern whether the predicted map was excessive or deficient, CC excels at providing a comprehensive evaluation of the entire map.

Earth Mover's Distance (EMD)

EMD is one of the evaluation metrics in similarity calculation methods for images that can consider spatial information. Specifically, EMD assesses the spatial distance between two probability distributions in a given region, and a higher evaluation value indicates a larger difference between the two distributions. Therefore, although it signifies that the two distributions are identical when EMD is 0, the EMD value increases when the predicted saliency map extends over a wide area. This indicates that EMD is a metric that strongly reacts to excessive predictions owing to its interpretation of imposing a strong penalty on false positives.

Similarity or Histogram Intersection (SIM)

SIM, a metric in image similarity calculation, derives evaluation values from the histograms of the predicted saliency map and the Ground Truth. Both maps are normalized beforehand. A value close to 1 indicates high similarity between the distributions, while a value near 0 indicates less overlap.

Kullback-Leibler Divergence (KL)

KL, an information-theoretic measure, evaluates the similarity of two probability distributions. The KL employed in salinecy map evaluation is asymmetric, with a lower value indicating that the predicted saliency map more accurately approximates the Ground Truth.

Information Gain (IG)

IG is an evaluation metric devised for saliency map prediction, drawing inspiration from information theory. In IG, it is assumed that the predicted saliency map follows the prob-

ability distribution, and the evaluation value is calculated after normalizing with considering center bias [37, 38]. This evaluation value reflects the prediction accuracy of the intrinsic saliency of the image without the effects of center bias.

2.3 Emotional Category Classification of Image

This section describes previous works related to emotional category classification methods. This thesis treats the emotional category classification for confirming the effectiveness of a uniform representation including visual contents and gaze data. Many studies on emotional category classification have focused on accurately capturing the relationship between images and emotional category, and they do not use any types of biological data. Therefore, this section introduces several representative references using only images as inputs.

Reference [39]

This study proposes a dataset constructed for the purpose of classifying emotional categories in images. Furthermore, this study validates the effectiveness of the dataset by conducting emotional category classification based on object-based features obtained from images, considering the report that emotions recalled by individuals during image viewing are related to objects in images.

References [40, 41]

These studies present emotional category classification methods using CNNs. The approach utilizes CNNs capable of accurately recognizing objects within images, leading to precise emotional category classification.

References [42, 43]

Zhao *et al.* directed their attention to identifying the common factors linking emotion features and visual features to predict emotion distribution. Under the assumption that a multitude of images are predetermined for the emotion distribution, they derived emotion features by considering the emotional distribution inherent in images.

References [44]

Lee *et al.* pay attention to semantic information obtained from objects for the performance

improvement, and the pre-trained word embedding succeed in extracting such information through the introduction of the attention mechanisms.

2.4 Multi-modal Human Emotion Recognition

This section shows the previous works related to multi-modal human emotion recognition. Although there are various problem settings when recognizing human emotions, this thesis focuses on human emotions recalled when viewing images.

Reference [45]

This study employs Deep Canonical Correlation Analysis (Deep CCA) [46] as the integration method. The study cohesively learns networks based on Deep CCA for gaze data and brain activity data, followed by the integration of features obtained from each network using a straightforward decision-label fusion.

Reference [47]

In this work, a more accurate emotion estimation is achieved by generating a mask image by transparently overlaying the image with the gazed region as a mask. Hand-crafted image features are then extracted from this mask image, resulting in higher accuracy in emotion prediction compared to using only the image.

References [48]

This study employs Bi-modal Deep Autoencoder (BDAE) [49] as the integration method. BDAE learns the relationship between the several types of biological information and constructs the common space. Given the gaze and brain activity data, the features calculated from each data are transformed into the common space for feature integration. Finally, the simple machine learning model with the integrated features outputs the final recognition results.

References [50]

In this study, Bi-modal Long-short Term Memory (BLSTM) [51] is used for feature integration with the consideration of the temporal changes in biological data. Given the input

data, the features calculated at each timestep are subsequently input into BLSTM. In this way, the temporal changes in biological data are considered when integrating features.

References [52]

In this study, a two-stream heterogeneous Graph Recurrent Neural Network is utilized to leverage the complementarity inherent in spatial-spectral-temporal domain features. The framework consists of multiple modules, including the graph transformer network designed to handle heterogeneity, the graph convolutional network aimed at capturing correlations, and the gated recurrent unit employed for analyzing dependencies in the temporal or spectral domain.

2.5 Problems to be Solved in this Thesis

This section clarifies the problems to be solved in this thesis. As summarized above, previous works have adopted general machine learning models such as multi-task CNN and Deep CCA for personalized prediction of human perception toward visual stimuli. Such general machine learning models may suffer from the inherent properties of biological data such as a small amount of training data, complex relationship between contents of visual stimuli and biological data, and the insufficient integration of several types of biological data. Therefore, there are great demands for rethinking the machine learning models suitable for biological data. Specifically, this thesis focuses on three situations, personalized saliency prediction, emotional category classification of images, and multi-modal human emotion recognition. Besides, for these situations, several machine learning models are newly constructed specific to each situation. Figure 2.1 shows a research map of related works that summarizes the above.

2.6 Conclusions

This chapter has summarized the previous works related to this thesis, including saliency prediction, emotional category classification of images, and multi-modal human emotion recognition. Furthermore, this chapter has clarified the problems to be solved in this thesis.

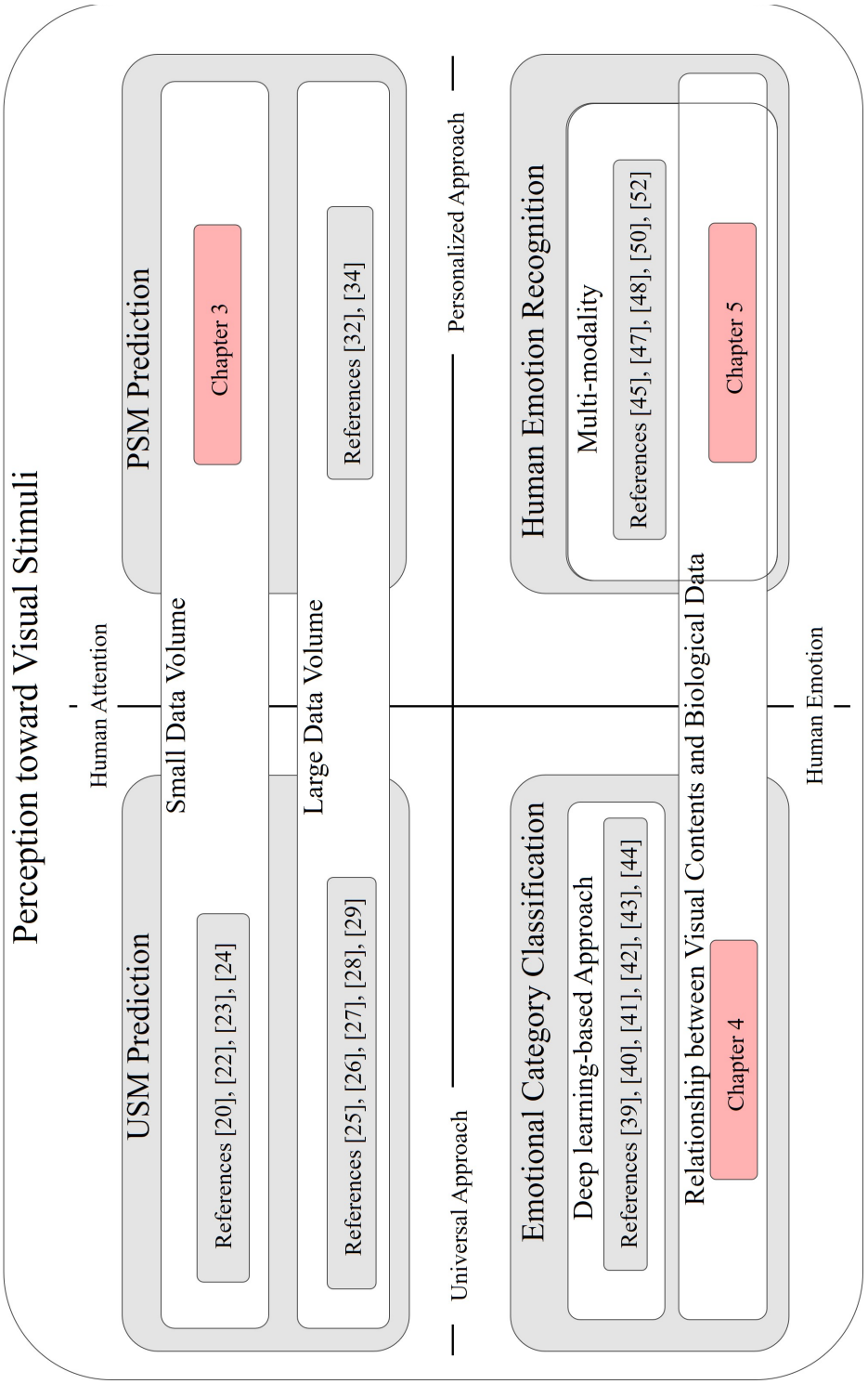

Figure 2.1: Research map of related works and position of each chapter in this thesis.

Chapter 3

Personalized Saliency Prediction

Visual saliency represents a distinctive subjective perceptual mechanism that enables humans to promptly identify crucial information in a complex environment. In the field of image processing, several works such as salient object detection [53, 54] and visual saliency prediction [20, 22, 23, 29] aim to implement the human visual attention mechanism on computers. Then, Universal Saliency Maps (USMs), which highlight salient regions in images, find diverse applications such as image re-targeting [55, 56], image enhancement [57, 58], and image compression [59, 60]. To model human instinctive perception, USMs are computed to emphasize regions that garner more attention than their surroundings [20]. In contrast, Personalized Saliency Maps (PSMs) consider individual visual attention variations, linked to personalized preferences [34, 61, 62], attracting considerable attention for PSM prediction [63, 64].

Gaze patterns may exhibit variations even when presented with the same image across different individuals, adding to the intricacy of extracting patterns for PSM prediction. In order to accomplish PSM prediction, it is imperative to collect data from individuals to analyze their gaze patterns and tendencies. Previous research [34] involved the acquisition of gaze data from 30 individuals exposed to diverse images, endeavoring to predict PSMs using such an extensive dataset. The PSM prediction approach relies on a multi-task Convolutional Neural Network (multi-task CNN) [33]. Trained on a substantial amount of individual gaze data, this network is capable of simultaneously predicting PSMs for multiple individuals. However, applying a multi-task CNN to a new individual without sufficient gaze data necessitates acquiring extensive data, posing a significant burden on the individual. Consequently, there is a need to develop a PSM

prediction method trainable with a limited amount of gaze data. As mentioned earlier, extracting gaze patterns and tendencies from a limited amount of training data poses a challenging task.

Chapter 3.1

Few-shot Personalized Saliency Prediction Based on Adaptive Image Selection Considering Object and Visual Attention

3.1.1 Introduction

The previous study [65] unveiled that utilizing gaze data from individuals who observe image regions akin to the target individual proves effective in predicting PSM. Under the assumption made in the previous study that individuals with similar characteristics have previously viewed the new image, it becomes viable to utilize the actual gaze data from these individuals. Nevertheless, as a new image may not always be gazed by other individuals, a more practical approach involves predicting the PSM of the target individual using PSMs predicted for other individuals with similar gaze patterns. The construction of such a method presents a difficult yet essential challenge.

To predict PSMs for the new target individual, the analysis of multiple images is needed in order to identify individuals with gaze patterns similar to those of the target individual. Before this process, it is crucial to select images from the extensive dataset to compute individual similarities between the target individual and those included in the dataset. Nevertheless, the reliability of the computed individual similarities is compromised if the chosen images exhibit high visual similarities to each other. Achieving robust PSM prediction with a reduced number of selected images necessitates an adaptive image selection scheme to address this concern. Two key as-

pects are particularly emphasized: 1) diversity of images and 2) variance of PSMs. Given the considerable diversity in images within the dataset, the selection of images while maintaining diversity becomes pivotal. Additionally, the variance of PSMs among individuals in the dataset should be high, as regions commonly viewed or ignored by many individuals can be effectively represented by USMs. The introduction of an adaptive image selection scheme that focuses on these aspects is anticipated to yield precise PSM prediction for the new target individual.

This chapter presents Few-shot PSM Prediction (FPSP) based on adaptive image selection (AIS) considering object and visual attention. The illustration in Fig. 3.1.1 delineates the problem under examination. To begin, a multi-task CNN is constructed and trained using the PSM dataset to predict PSMs for individuals within the dataset [33]. Subsequently, the similarity between individuals is computed using chosen images from the PSM dataset, selected by AIS to emphasize image diversity and PSM variance. To ensure the diversity of selected images, AIS concentrates on the types of objects found in training images within the PSM dataset, employing a object detection approach. Identified objects with substantial PSM variances are pinpointed, enabling the adaptive selection of images containing such objects, as illustrated in the orange region of Fig. 3.1.1. Finally, the FPSP for a target image concerning a new individual is executed based on the individual similarity and PSM predictions derived from the multi-task CNN trained on individuals in the PSM dataset. In this way, FPSP utilizing AIS for the new individual can be achieved with high accuracy even with a limited training dataset.

Figure 3.1.1: Problem setting. The purpose of this chapter is PSM prediction of a target individual for images absent from the training dataset. For predicting a PSM for a target image, the target individual needs to view only some images, which have been viewed by individuals included in the training PSM dataset.

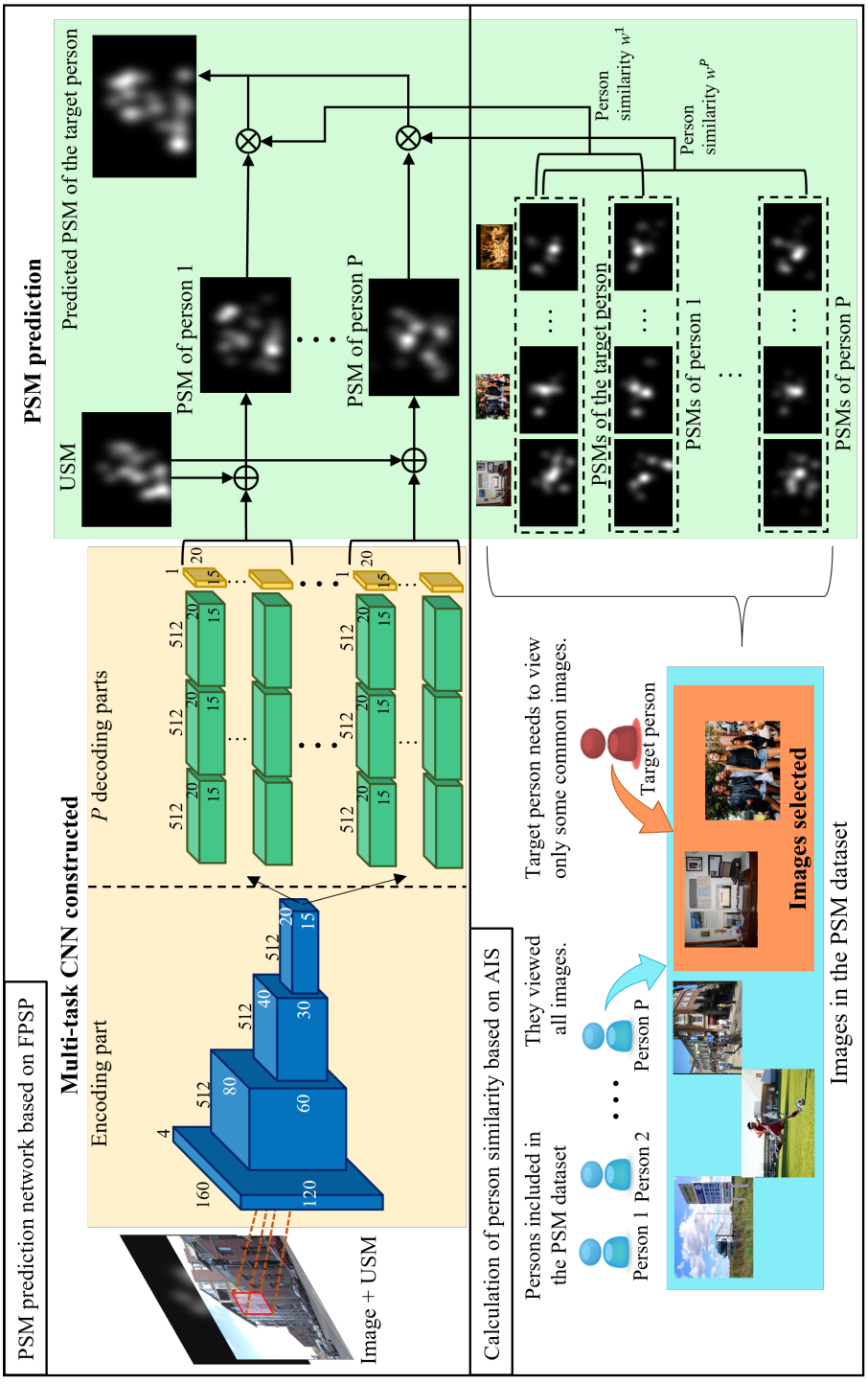

Figure 3.1.2: Outline of FPSP based on AIS. The top row illustrates the FPSP process, while the bottom row depicts the approach used in computing individual similarities through AIS.

3.1.2 Proposed PSM Prediction

This section describes our proposed method as illustrated in Fig. 3.1.2. Our approach involves training a multi-task CNN to predict PSM for individuals included in the large-scale dataset. Subsequently, image selection based on AIS is performed to select used images. The target individual is then required to view only the selected images for their PSM prediction. Finally, the PSM of the target image for the target individual is predicted by leveraging the PSMs predicted for similar individuals.

3.1.2.1 Multi-task CNN Construction

The multi-task CNN is configured to compute P PSMs, where P represents the number of individuals. In our proposed method, the input data, which is used for training the multi-task CNN, comprises images $\boldsymbol{X}_n \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ($n = 1, 2, \ldots, N$; N denoting the number of training images, $d_1 \times d_2$ indicating the number of pixels, and d_3 representing the number of color channels) and USMs $\boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n) \in \mathbb{R}^{d_1 \times d_2}$. The USM denotes the region where multiple individuals view. In our approach, the USM $\boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)$ is obtained through the USM prediction method. For the given PSMs $\boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{X}_n) \in \mathbb{R}^{d_1 \times d_2}$ ($p = 1, 2, ..., P$) for P individuals, where $\boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{X}_n)$ is derived from the gaze data of the pth individual for the image $\boldsymbol{X}_n$, we compute a difference map $\boldsymbol{\Delta}(p, \boldsymbol{X}_n)$ between the PSM of the individual p and the USM in accordance with [34] as follows:

$$\boldsymbol{\Delta}(p, \boldsymbol{X}_n) = \boldsymbol{S}^{\mathrm{PSM}}(p, \boldsymbol{X}_n) - \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n). \tag{3.1.1}$$

The multi-task CNN consists of one encoder and P decoders and each comprising three layers. The output layer yields P results of $\boldsymbol{\Delta}(p, \boldsymbol{X}_n)$. The model architectures of the multi-task CNN are illustrated in Fig. 3.1.3. Additionally, the training of the multi-task CNN involves minimizing the following loss function:

$$\sum_{l=1}^{3} \sum_{p=1}^{P} \sum_{n=1}^{N} \|\hat{\boldsymbol{\Delta}}_l(p, \boldsymbol{X}_n, \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)) - \boldsymbol{\Delta}(p, \boldsymbol{X}_n)\|_F^2, \tag{3.1.2}$$

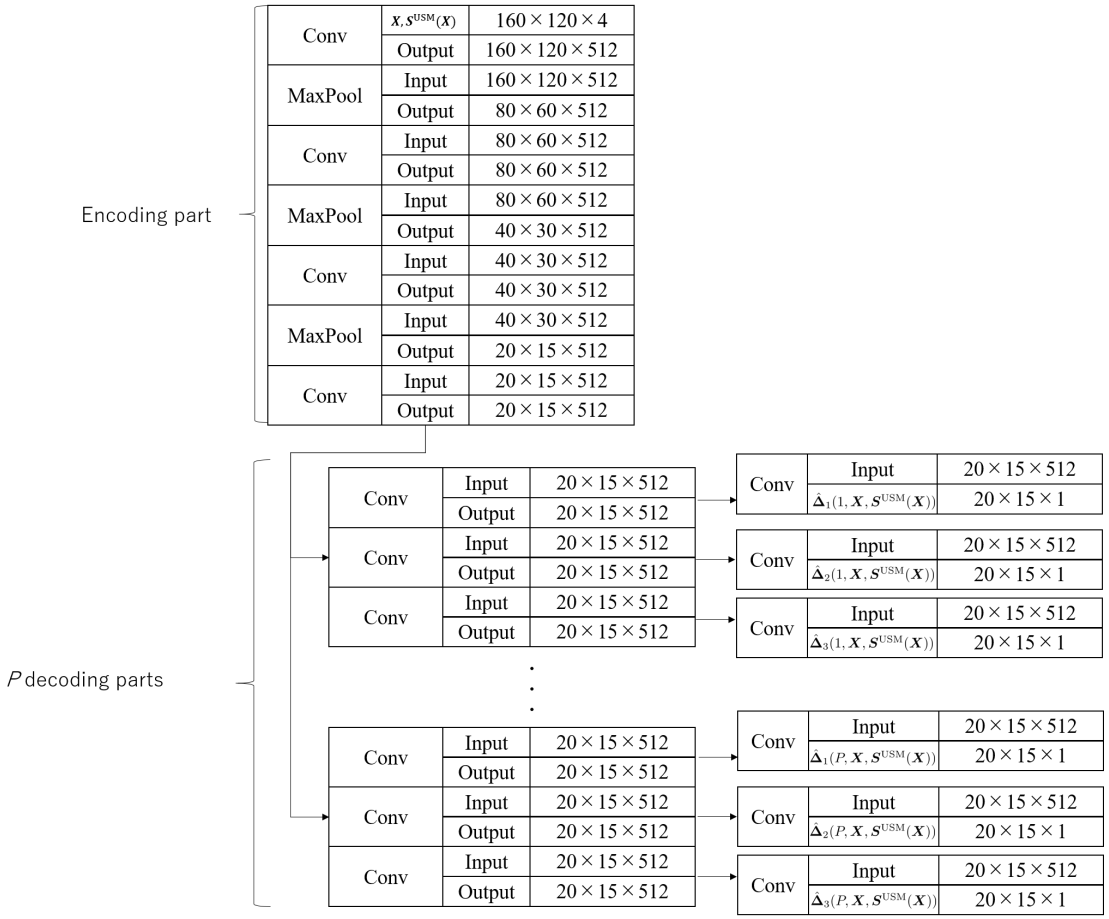

Figure 3.1.3: The model architecture of the multi-task CNN employed in our approach. This figure uses "Conv" and "MaxPool" for indicating the application of a convolution and max-pooling layer, respectively.

where $\hat{\mathbf{\Delta}}_l(\cdot)$ represents a function for calculating the difference map and employs a 1×1 convolution layer on the outputs obtained from the lth decoding layer, and $\|\cdot\|_F^2$ denotes the operator for the squared two-order Frobenius norm. When given a new image $\boldsymbol{X}^{\mathrm{tgt}}$, the prediction of the PSM for individual p is calculated using the trained network as the following expression:

$$\boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}^{\mathrm{tgt}}) = \hat{\mathbf{\Delta}}_3(p, \boldsymbol{X}^{\mathrm{tgt}}, \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}^{\mathrm{tgt}})) + \boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}^{\mathrm{tgt}}). \tag{3.1.3}$$

In this way, the multi-task CNN allows the prediction of PSMs for multiple individuals using a single model.

3.1.2.2 Adaptive Image Selection

The purpose of AIS is to reduce the number of images gazed at by the target individual for predicting their PSMs. For a new individual p^{new} not belonging to the individuals in the training data for the multi-task CNN, the multi-task CNN cannot learn their PSMs since they do not view all images in the training data for the multi-task CNN. To address this, we obtain a limited amount of seed PSMs for images from the target individual. The selection of images viewed by the target individual is crucial for reducing their burden, as the influence of each image on training is substantial. Image diversity significantly relies on the selection scheme, even though the PSM dataset [33] inherently possesses diverse images. Thus, we propose an image selection method that maintains diversity by considering object types and PSM variances as illustrated in Fig. 3.1.4. To maximize object variety in the selected images, we apply the object detection method [66] to images that are used for training the multi-task CNN. Additionally, we use gaze data obtained from individuals for pre-selected images to calculate the PSM variances. In AIS, image selection is based on the detected objects and their associated PSM variances. Specifically, we choose objects with high PSM variances, as those with low variances are expected to be represented by USMs. Finally, we select images containing diverse objects with high PSM variances. In the first step of the AIS procedure, objects $\boldsymbol{O}_{(n,m)}$ ($m = 1, \ldots, M$; M denoting the types of objects across all images) are detected in the images using the object detection methods. Detected objects are represented by bounding boxes with dimensions $d^h_{(n,m)} \times d^w_{(n,m)}$. Subsequently, we calculate the object variance $v_{(n,m)}$ as follows:

$$v_{(n,m)} = \frac{1}{d^h_{(n,m)} \times d^w_{(n,m)}} \sum_{j=1}^{d^h_{(n,m)}} \sum_{k=1}^{d^w_{(n,m)}} \frac{1}{P} \sum_{p=1}^{P} \left\{ \boldsymbol{S}^{\text{PSM}}(p, \boldsymbol{O}_{(n,m)})_{(j,k)} - \bar{\boldsymbol{S}}^{\text{PSM}}(\boldsymbol{O}_{(n,m)})_{(j,k)} \right\}^2, \tag{3.1.4}$$

$$\bar{\boldsymbol{S}}^{\text{PSM}}(\boldsymbol{O}_{(n,m)})_{(j,k)} = \frac{1}{P} \sum_{p=1}^{P} \boldsymbol{S}^{\text{PSM}}(p, \boldsymbol{O}_{(n,m)})_{(j,k)}, \tag{3.1.5}$$

where $\boldsymbol{S}^{\text{PSM}}(p, \boldsymbol{O}_{(n,m)})$ denotes the PSM of individual p corresponding to the object $\boldsymbol{O}_{(n,m)}$, and (j, k) denotes the pixel location. It is important to note that we consider $v_{(n,m)} = 0$ if image $\boldsymbol{X}_n$ does not contain the mth object, and the highest $v_{(n,m)}$ is selected if image $\boldsymbol{X}_n$ includes multiple mth objects. To conduct our image selection, we compute the sum of variances, $\bar{v}_n$, for PSMs

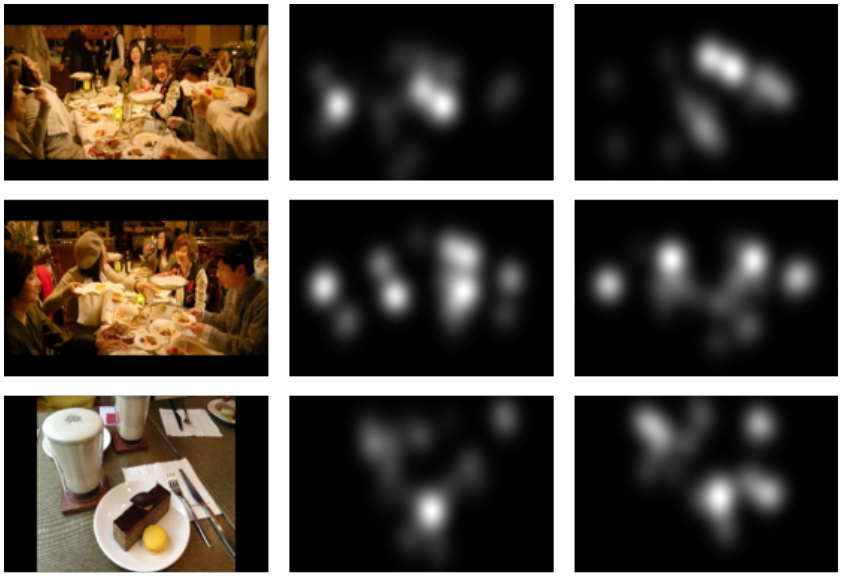

Figure 3.1.4: Examples are presented to describe the diversity of images and the PSM variance. The images in the first and second rows exhibit visual similarities, which prompts AIS to choose either one. In contrast, the image in the third row is bypassed by AIS due to the resemblance in PSMs between persons.

associated with each image using the following formula:

$$\bar{v}_n = \sum_{m=1}^{M} v_{(n,m)}. \tag{3.1.6}$$

In conclusion, C images with the highest values in Eq. (3.1.6) are chosen. Given that human visual attention is influenced by objects, our approach, which explicitly considers this connection, realizes a simple but effective approach for preserving the image diversity and the PSM variance. Therefore, AIS, with its emphasis on combining object detection and visual attention, is a potent strategy.

3.1.2.3 Individual Similarity-based FPSP

The proposed method predicts the PSM of the new individual p^{new} based on the PSMs predicted for the similar individuals by the multi-task CNN. In the first step of PSM prediction of the new individual, $\boldsymbol{S}^{\text{out}}(p, \boldsymbol{X}_c^{\text{sel}})$ is calculated by feeding the target image into the multi-task

CNN with Eq. (3.1.3). Here, $\boldsymbol{X}_c^{\mathrm{sel}}$ $(c = 1, 2, \ldots, C)$ represents the C images selected based on AIS. Subsequently, by using the predicted PSMs $\boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}_c^{\mathrm{sel}})$, we compute the correlation as a similarity score β^p between the target individual p^{new} and the individual p as follows:

$$\beta^p = \frac{1}{C} \sum_{c=1}^{C} \mathrm{corr}\left(\boldsymbol{S}^{\mathrm{PSM}}(p^{\mathrm{new}}, \boldsymbol{X}_c^{\mathrm{sel}}), \boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}_c^{\mathrm{sel}})\right), \tag{3.1.7}$$

where $\mathrm{corr}(\cdot, \cdot)$ calculates the correlation coefficient, and $\boldsymbol{S}^{\mathrm{PSM}}(p^{\mathrm{new}}, \boldsymbol{X}_c^{\mathrm{sel}})$ is derived using the gaze data of the target individual p^{new}. This implies that the new individual p^{new} is required to view only the selected C images to acquire gaze data for computing the PSM $\boldsymbol{S}^{\mathrm{PSM}}(p^{\mathrm{new}}, \boldsymbol{X}_c^{\mathrm{sel}})$. Subsequently, to mitigate the influence from dissimilar individuals, we exclusively choose similar individuals based on the selection coefficient a^p, which is determined as follows:

$$a^p = \begin{cases} 1 & (\beta^p > \tau) \\ 0 & (\text{otherwise}), \end{cases} \tag{3.1.8}$$

where τ denotes a predetermined threshold value. Subsequently, leveraging the similarity score and the selection coefficient, the individual similarity between the individual p and the new individual p^{new} is computed in the following equation:

$$w^p = \frac{a^p \beta^p}{\sum_{p'} a^{p'} \beta^{p'}}. \tag{3.1.9}$$

By employing the individual similarities w^p and the predicted PSMs of similar individuals from the multi-task CNN, we can straightforwardly predict the PSM $\boldsymbol{S}^{\mathrm{FPSP}}(p^{\mathrm{new}}, \boldsymbol{X}^{\mathrm{tgt}})$ for the new individual p^{new} and the target image $\boldsymbol{X}^{\mathrm{tgt}}$ as follows:

$$\boldsymbol{S}^{\mathrm{FPSP}}(p^{\mathrm{new}}, \boldsymbol{X}^{\mathrm{tgt}}) = \sum_{p=1}^{P} w^p \boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X}^{\mathrm{tgt}}). \tag{3.1.10}$$

In this way, leveraging the individual similarity w^p, the proposed method facilitates the prediction of the PSM for the new individual with a limited amount of training gaze data.

3.1.3 Experiments

3.1.3.1 Settings

This experiment was conducted on the PSM dataset [33], consisting of 1600 images, was employed, along with associated gaze data for 30 individuals with either corrected-to-normal or normal vision. Gaze data were recorded as each individual viewed each image for three seconds under free-viewing conditions. PSMs were computed based on gaze data following the methodology outlined in [67] and were employed as the Ground Truth (GT). For the experiment, 500 images were randomly selected as test images, while the remaining 1100 images were used for training. Additionally, we varied the number of selected training images, which is denoted as C, within $\{10, 20, \ldots, 100\}$. In this experiment, we randomly selected 10 individuals as new targets, employing the remaining 20 individuals for training the multi-task CNN. The multi-task CNN was optimized using stochastic gradient descent [68], with learning rate, mini-batch size, the number of iterations set, and momentum at 0.00003, 9, 1000, and 0.9, respectively. The threshold value τ was experimentally set to 0.7. In the proposed method, $\boldsymbol{S}^{\mathrm{USM}}(\boldsymbol{X}_n)$ could be computed as the average of PSMs from the training set of 20 individuals which were used for training the multi-task CNN.

To assess the effectiveness of the proposed method encompassing the image selection scheme, we conducted qualitative and quantitative evaluations. Quantitatively, we measured the difference between predicted PSMs and their GT using CC, KLdiv, and Sim presented in Section 2.2.3. Additionally, two types of comparative experiments were performed. Initially, we directed our attention to assessing the efficacy of our approach with a limited quantity of gaze data. To gauge the performance, we compared our method with the following existing methods, signature [23], GBVS [22], Itti [20], and SalGAN [29], which are USM prediction methods from the MIT saliency benchmark [69]. SalGAN is trained with the SALICON dataset [70] Additionally, we compared our method with two PSM prediction methods as follows:

- PSM prediction using visual similarities (Baseline1) [71]
- PSM prediction using visual similarities and spatial information (Baseline2) [72]

It is noteworthy that we trained the aforementioned comparative methods with images chosen by AIS, assuming that the target individual views only these images.

In the second comparative analysis, aimed at underscoring the efficacy of our image selection methodology, we conducted a comparison with the following image selection techniques:

- Image selection based on visual features (ISVF)
  Images were chosen by evaluating the dissimilarity of visual features to other images, utilizing outputs from the final pooling layer of the pre-trained DenseNet201 [73] as visual features.

- Image selection focusing on the variance of PSMs (ISPSM)
  Selection involved choosing images with a high variance in PSMs based on the PSMs of the individuals used for training the multi-task CNN.

3.1.3.2 Results and Discussions

Experimental results are illustrated in Figs. 3.1.5 - 3.1.12, and Table 3.1.1 presents evaluation scores. Figure 3.1.5 displays the predicted outcomes for one individual for demonstrating that the FPSP method excels in predicting a PSM that closely aligns with the GT compared to all other PSMs predicted by the comparative methods. The average results are presented in Table 3.1.1, clearly indicating that FPSP based on AIS stands out as the most effective approach for PSM prediction across all evaluation indices. In this way, through the comparison of averages, we affirm the notable efficacy of the proposed method.

The outcomes predicted by FPSP based on AIS and the USM prediction methods are illustrated for individual participants in Figs. 3.1.6 - 3.1.8. It is noteworthy that we label the 10 target individuals as Pars 1-10 in these figures. These visuals demonstrate that FPSP successfully achieves individual-specific predictions for the majority of participants, outperforming the USM prediction methods. This observation confirms the efficacy of constructing a personalized prediction model for each individual. Additionally, Figs. 3.1.6 - 3.1.11 present the outcomes for each participant obtained by the PSM prediction methods and FPSP based on AIS, demonstrating superior results compared to alternative PSM prediction methods. Consequently, FPSP

demonstrates more precise predictions than baseline PSM prediction methods, validating its effectiveness in the first experiment.

Moving on to the second experiment, we delve into the distinctions among AIS, ISVF, and ISPSM. Analyzing the baselines in Table 3.1.1, it is evident that AIS emerges as the most impactful image selection method. Additionally, in Fig. 3.1.12, the performance of FPSP is depicted with changes in the number of training images, selected by AIS, ISVF, and ISPSM to compute individual similarity. In essence, FPSP based on AIS accurately predicts the PSMs of the target individuals with just 10 images from the PSM dataset. Hence, our image selection method, AIS, is evidently effective for FPSP. Consequently, the experimental results confirm the robustness and efficacy of FPSP based on AIS.

In summary, our discussions affirm the effectiveness of the proposed PSM prediction method, FPSP, in Fig. 3.1.5 and Table 3.1.1, considering both qualitative and quantitative evaluations. Furthermore, the comparisons of FPSP with USM prediction methods and baseline PSM prediction methods for each individual in Figs. 3.1.6 - 3.1.11 validate its capability to achieve accurate predictions for individuals. Finally, the robustness and efficacy of AIS for FPSP are substantiated by Fig. 3.1.12. In conclusion, FPSP based on AIS stands out as a method that enables accurate predictions with a limited number of training images, thereby alleviating the burden on individuals for obtaining gaze data for PSM prediction.

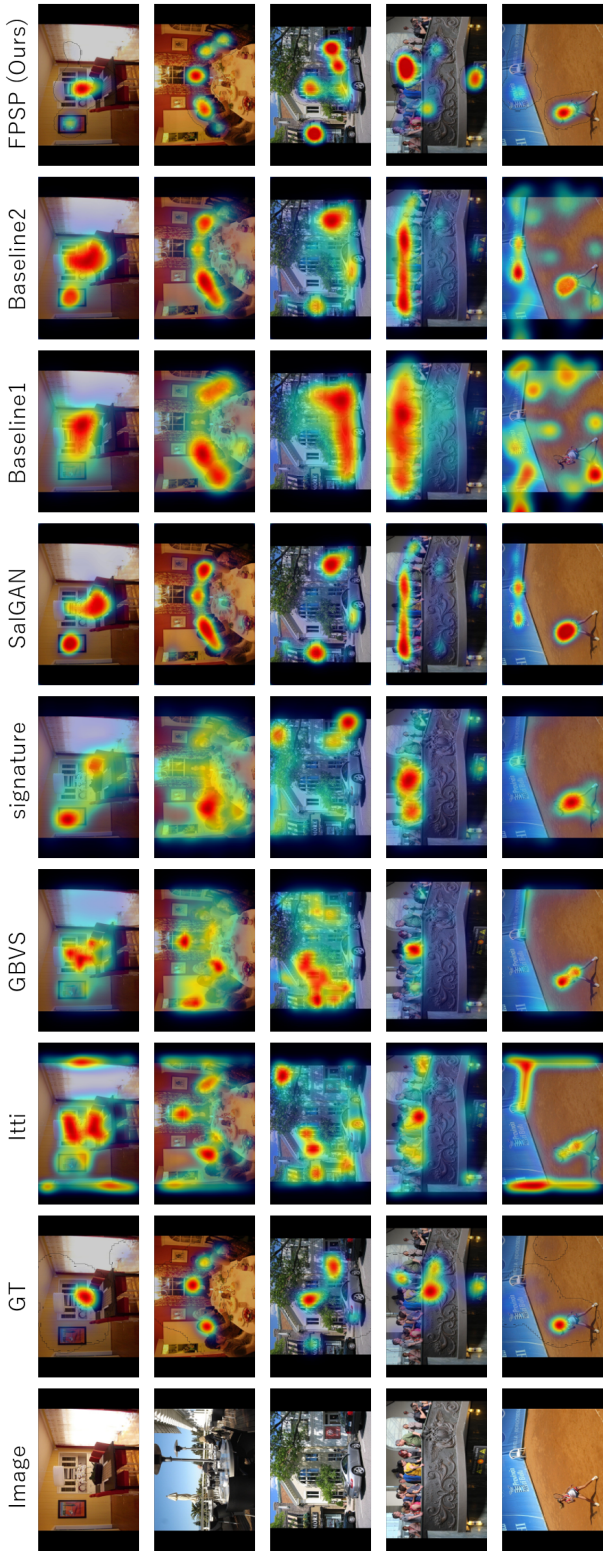

Figure 3.1.5: Qualitative outcomes for one individual predicted by the proposed and comparative methods. In this illustration, the training images for Baselines 1 and 2, along with FPSP, were selected using AIS.

Table 3.1.1: Performance comparison across various evaluation indices. The symbol (↑) indicates that a higher index corresponds to improved performance, while the symbol (↓) indicates that a lower index reflects improved performance. It is important to mention that 100 (=C) selected images were utilized for training in Baselines 1 and 2, as well as the proposed method. The use of bold font signifies the highest value within its respective evaluation index.

Methods	**CC↑**	**Sim↑**	**KLdiv↓**
Itti	0.3218	0.3911	9.0397
GBVS	0.4367	0.4474	6.8923
signature	0.4126	0.4122	8.0410
SalGAN	0.6345	0.5689	3.5597
Baseline1 based on ISVF	0.0953	0.3140	11.029
Baseline1 based on ISPSM	0.0762	0.3100	11.161
Baseline1 based on AIS	0.4013	0.4165	7.641
Baseline2 based on ISVF	0.4842	0.4274	4.014
Baseline2 based on ISPSM	0.4761	0.4170	3.057
Baseline2 based on AIS	0.5972	0.5032	4.133
FPSP based on AIS (Ours)	**0.7845**	**0.6557**	**1.083**

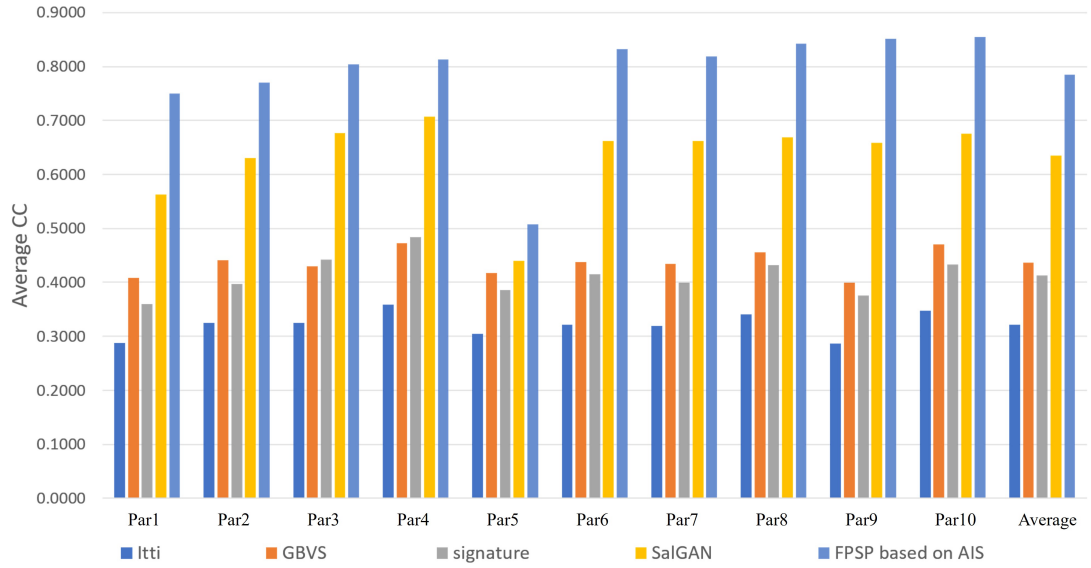

Figure 3.1.6: Average CC (↑) for each target individual for the proposed method and USM prediction methods.

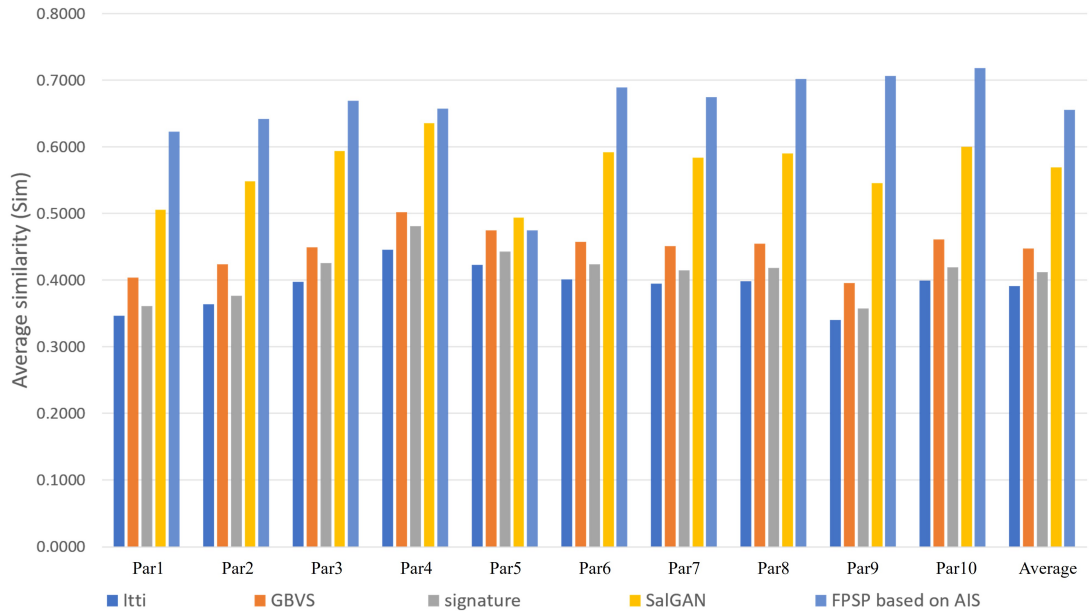

Figure 3.1.7: Average Sim (↑) for each target individual for the proposed method and USM prediction methods.

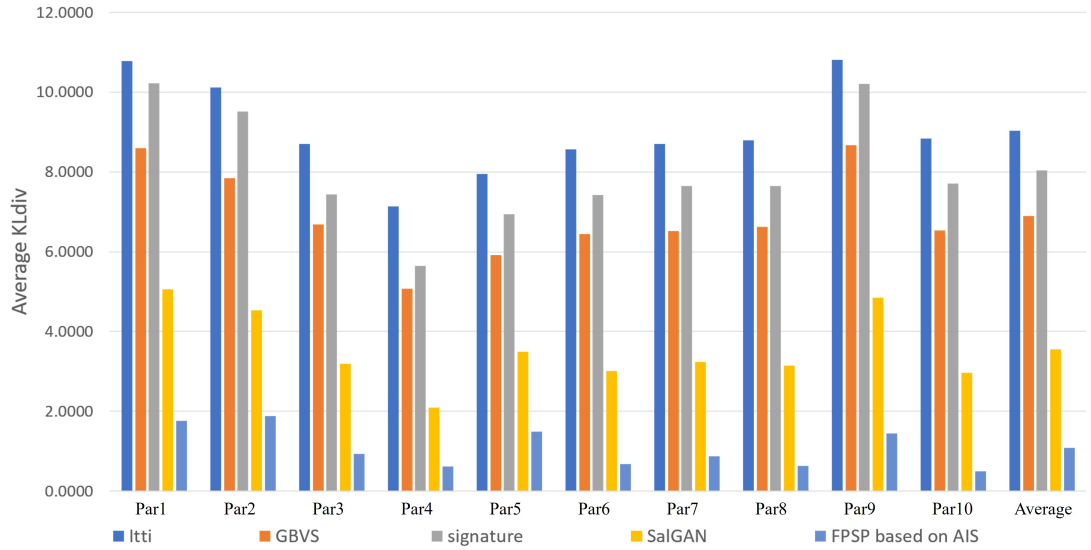

Figure 3.1.8: Average KLdiv (↓) for each target individual for the proposed method and USM prediction methods.

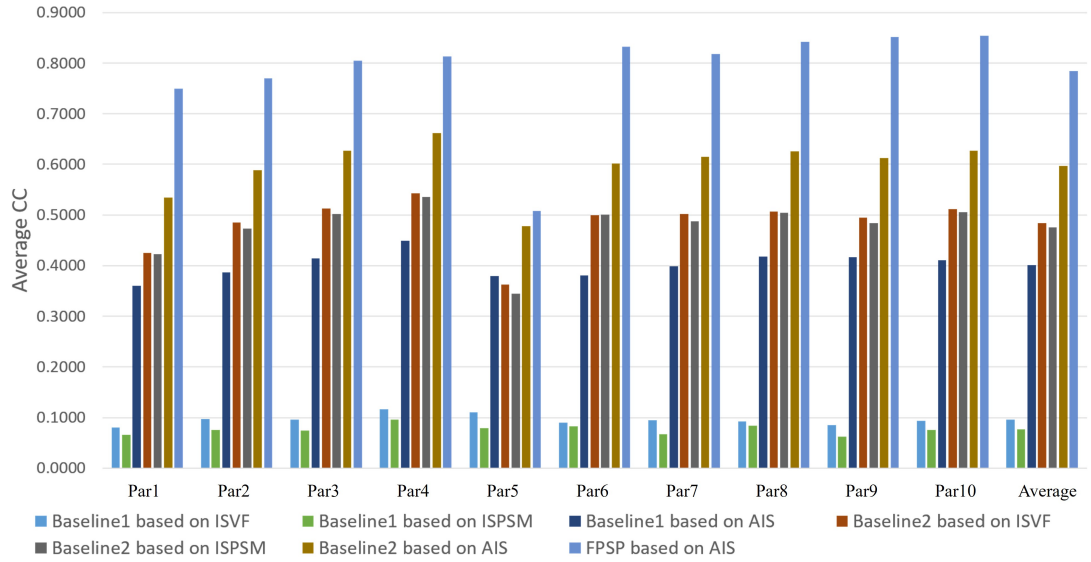

Figure 3.1.9: Average CC (↑) for each target individual for the proposed method and other PSM prediction methods.

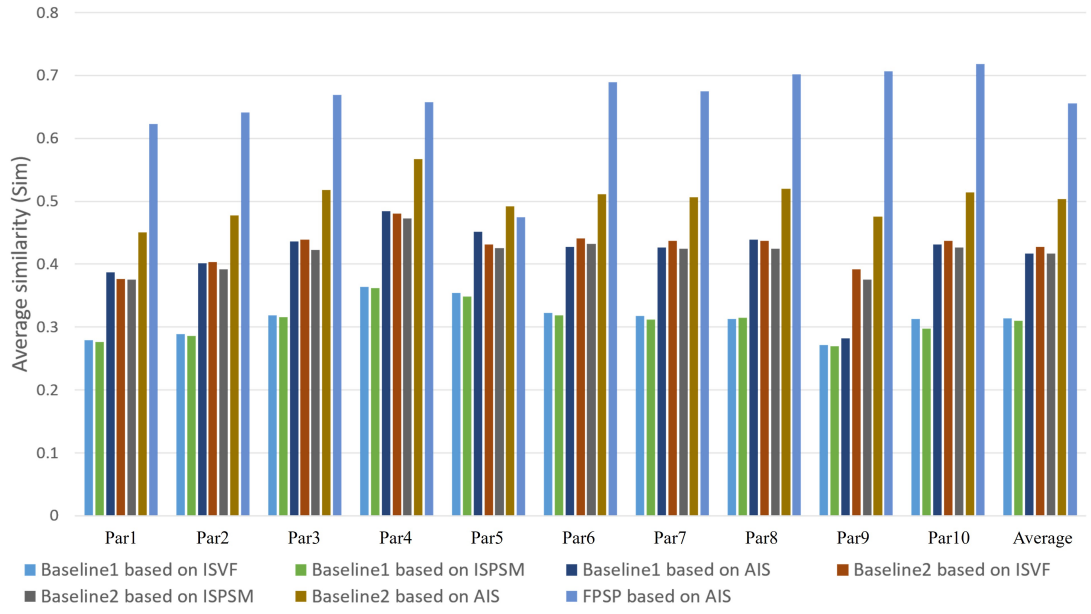

Figure 3.1.10: Average Sim (↑) for each target individual for the proposed method and other PSM prediction methods.

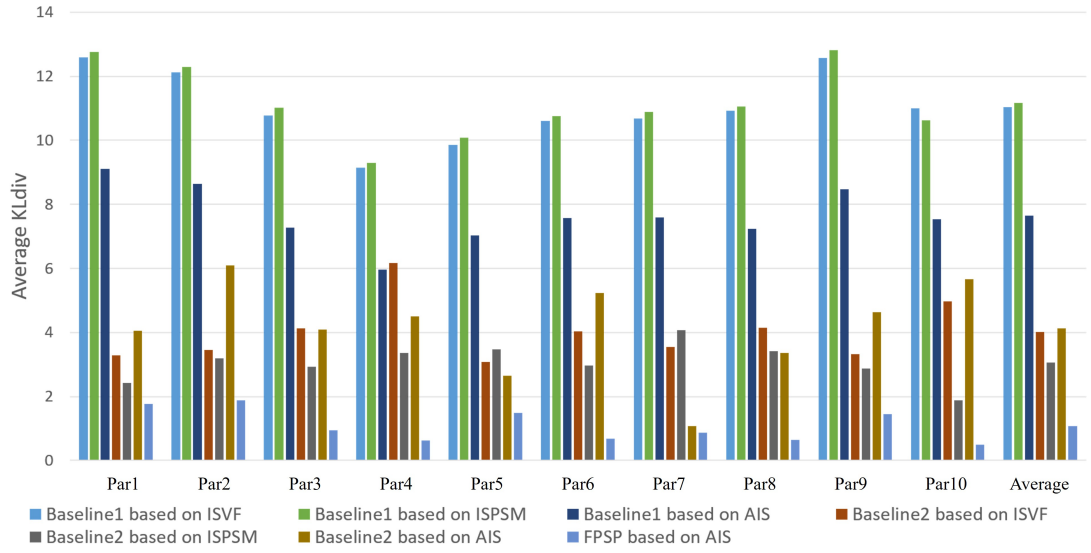

Figure 3.1.11: Average KLdiv (↓) for each target individual for the proposed method and other PSM prediction methods.

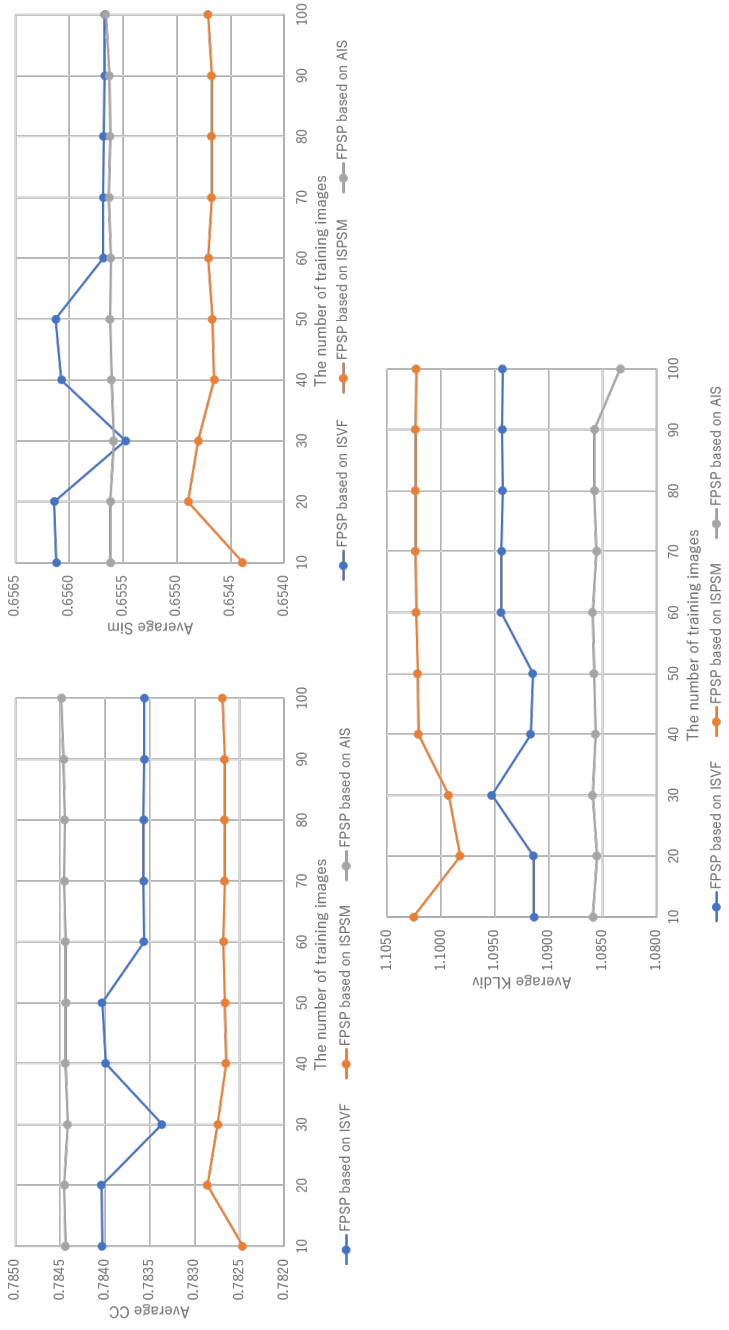

Figure 3.1.12: The prediction performance under varying numbers of training images. The resilience of FPSP based on AIS is demonstrated.

3.1.4 Conclusions

This chapter presents a approach for FPSP based on AIS, for taking into account visual attention and objects. FPSP enhances the precise PSM prediction with a limited number of training images. Additionally, AIS contributes to reducing the number of images observed by a new individual. Consequently, FPSP with AIS achieves accurate predictions with a reduced number of training images, alleviating the burden on individuals in acquiring gaze data for PSM prediction.The efficacy of the proposed method is validated thorough the experiment.

Chapter 3.2

Few-shot Personalized Saliency Prediction Using Individual Similarity Based on Collaborative Multi-output Gaussian Process Regression

3.2.1 Introduction

In Chapter 3.1, the focus has been on predicting the PSM for the target individual through similarities invariant to target images. However, the connection between gaze data or PSMs and the visual stimuli of the image [61] implies that the similarity among individuals can fluctuate from one image to another. Therefore, to achieve precise PSM prediction for the target individual, similarity calculation is needed for each image and each individual. As previously noted, the amount of gaze data from the target individual is limited, and deterministic machine learning methods may lead to overfitting on the training data. In such circumstances, it becomes imperative to devise a probabilistic approach for PSM prediction, incorporating the calculation of similarity for each image and each individual.

This chapter introduces Few-shot PSM Prediction (FPSP) using individual similarity based on Gaussian process regression (GPR). To predict the PSM for the target individual, we adopt GPR with the predicted PSMs of other individuals and visual features. GPR is renowned for mitigating overfitting to training data by leveraging probabilistic validation. Additionally, the incorporation of visual features into the inputs facilitates the consideration of image variations. Managing the PSM, which encompasses multiple variables such as saliency values correspond-

ing to pixel values necessitates a Multi-output GPR (MOGP) model. Among various MOGP models, Collaborative MOGP (CoMOGP) [74] is reported as one of the most proficient and promising methods [75]. CoMOGP can collectively consider information on output relationships and the outputs themselves. Consequently, by inputting the predicted PSM and visual features into CoMOGP, the proposed method accommodates similarity considerations for each image and individual, leading to PSM prediction for the target individual. The contributions in this chapter are two-folds:

(i) The probabilistic regression model predicts the PSM of the target individual without succumbing to overfitting with limited training data.

(ii) CoMOGP represents similarities between individuals as weights of input PSMs and incorporates visual stimuli of images by utilizing visual features in the input.

In this way, FPSP, leveraging visual attention similarity based on the CoMOGP model is anticipated to achieve high prediction accuracy with a minimal amount of gaze data.

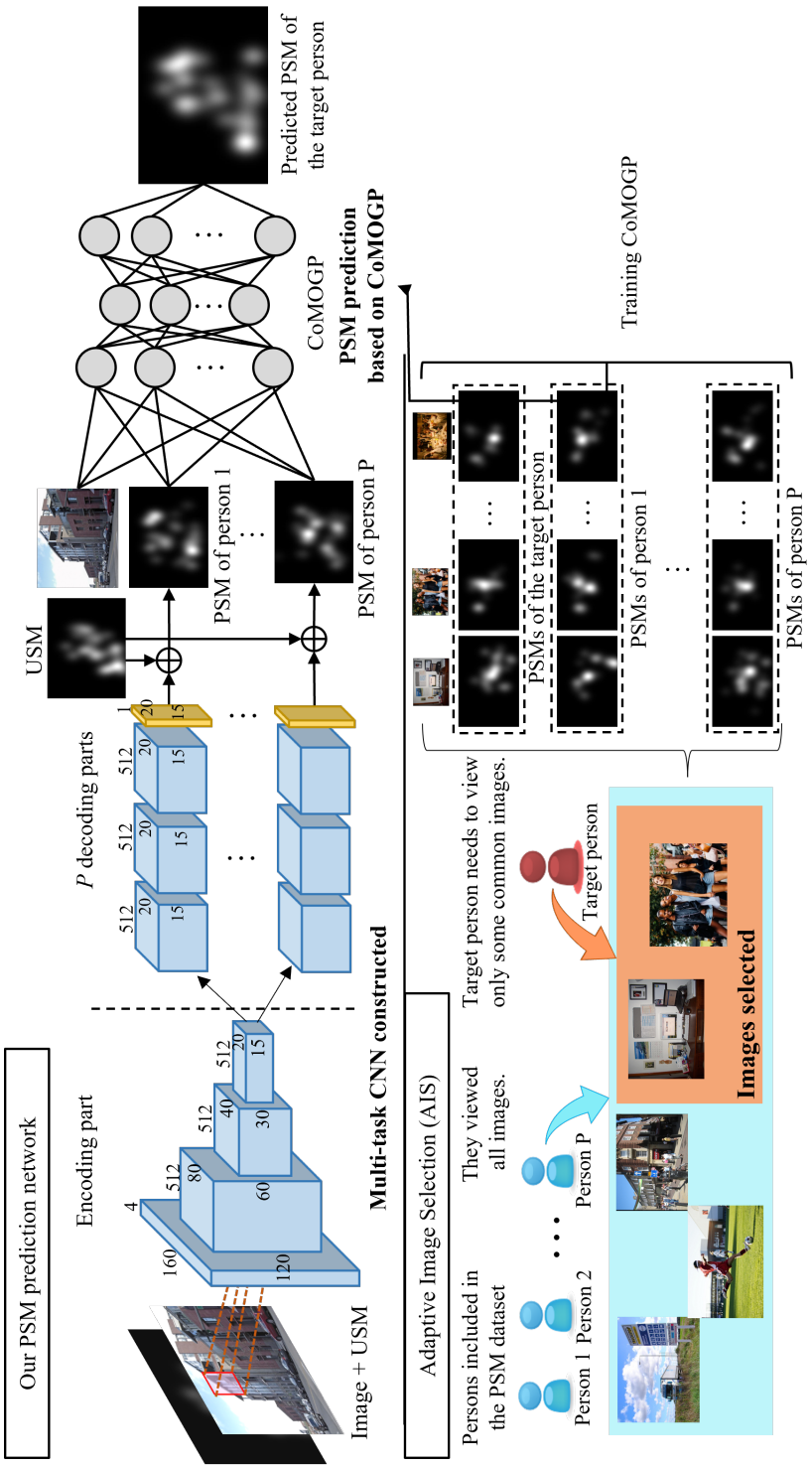

Figure 3.2.1: Flow of FPSP based on CoMOGP. Our approach involves the development of a multi-task CNN customized for individuals with a enough amount of training gaze data. Subsequently, the AIS process is applied to choose particular images for the target individual to view. Finally, leveraging visual features computed from images and PSMs predicted by the multi-task CNN, CoMOGP enables the prediction of the PSM for the target individual, even with a restricted amount of training gaze data.

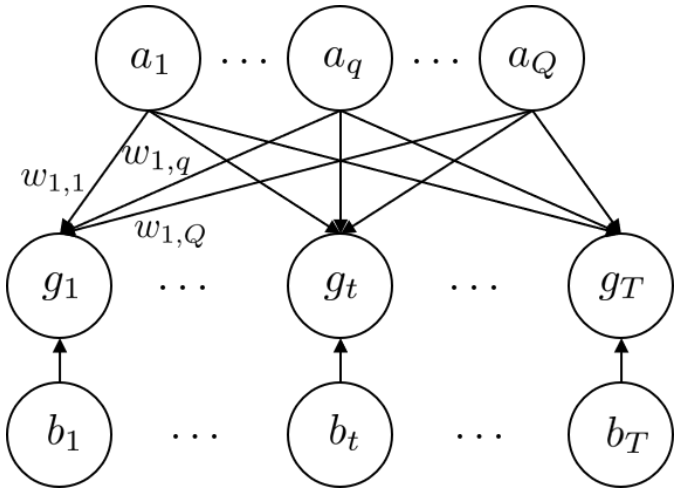

Figure 3.2.2: Graphical model of CoMOGP is presented, where a_q and b_t, presumed to conform to Gaussian processes, are computed utilizing the covariances of the input data. Moreover, $w_{t,q}$ denotes the weights associated with $a_q(\boldsymbol{I}_c)$ and $b_t(\boldsymbol{I}_c)$, while g_t represents the outputs.

3.2.2 Few-shot PSM Prediction Based on CoMOGP

This section provides a comprehensive understanding of FPSP based on CoMOGP as illustrated in Fig. 3.2.1. In the initial phase of the proposed method, the multi-task CNN undergoes training to predict the PSMs of individuals with a substantial amount of gaze data. Following this, we employ AIS to select specific images that the target individuals are need to view for PSM prediction. It is assumed that both the target and other individuals commonly view these chosen images. Lastly, CoMOGP utilizes PSMs predicted by the multi-task CNN and visual features computed from the target image to predict the PSMs of the target individual. In this section, CoMOGP is mainly explained, while multi-task CNN and AIS are explained in Sections 3.1.2.1 and 3.1.2.2, respectively.

3.2.2.1 FPSP Based on CoMOGP

Graphical model of CoMOGP is depicted in Fig. 3.2.2. The inputs for CoMOGP involve the visual features $\boldsymbol{f}_c \in \mathbb{R}^{d_f}$ calculated from the images $\boldsymbol{X}_c$ ($c = 1, 2, \ldots, C$; C representing the

number of selected images) chosen by AIS. The construction of inputs is as follows:

$$\boldsymbol{I}_c = [\mathrm{vec}(\boldsymbol{S}^{\mathrm{out}}(1, \boldsymbol{X}_c))^\top, \mathrm{vec}(\boldsymbol{S}^{\mathrm{out}}(2, \boldsymbol{X}_c))^\top, \ldots, \mathrm{vec}(\boldsymbol{S}^{\mathrm{out}}(P, \boldsymbol{X}_c))^\top, \boldsymbol{f}_c^\top]^\top, \tag{3.2.1}$$

where $\boldsymbol{S}^{\mathrm{out}}(p, \boldsymbol{X})$ represents the PSM of the image $\boldsymbol{X}$ for the individual p as predicted by the multi-task CNN. Note that $\mathrm{vec}(\cdot)$ denotes the vectorization process. To train CoMOGP, we formulate the inputs $\boldsymbol{I} = \{\boldsymbol{I}_c\}_{c=1}^C$ and the corresponding outputs $\boldsymbol{Y} = [\mathrm{vec}(\boldsymbol{S}(p^{\mathrm{tgt}}, \boldsymbol{X}_c))]_{c=1}^C \in \mathbb{R}^{CT}$, where p^{tgt} is the target individual assumed to view the selected images. In CoMOGP, its outputs are expressed as follows:

$$y_t(\boldsymbol{I}_c) = g_t(\boldsymbol{I}_c) + \sigma_t, \tag{3.2.2}$$

$$g_t(\boldsymbol{I}_c) = \sum_{q=1}^{Q} w_{t,q} a_q(\boldsymbol{I}_c) + b_t(\boldsymbol{I}_c), \tag{3.2.3}$$

where $w_{t,q}$ represents the weights of the latent values $a_q(\boldsymbol{I}_c)$ ($q = 1, 2, \ldots, Q$; Q being the number of latent values), and $b_t(\boldsymbol{I}_c)$ is features specific to $g_t(\boldsymbol{I}_c)$. It is noteworthy that $a_q(\boldsymbol{I}_c)$ and $b_t(\boldsymbol{I}_c)$ are assumed to follow Gaussian processes. Additionally, $y_t(\boldsymbol{I}_c)$ denotes the output values of the tth dimension of the outputs $\mathrm{vec}(\boldsymbol{S}(p^{\mathrm{tgt}}, \boldsymbol{X}_c)) \in \mathbb{R}^T$, and σ_t represents the Gaussian noise. The posterior distribution concerning the latent variables a_q and b_t is approximated through variational inference, and the evidence lower bound is optimized throughout the training process.

Given the target image $\boldsymbol{X}^{\mathrm{tgt}}$, the PSM $\boldsymbol{S}(p^{\mathrm{tgt}}, \boldsymbol{X}^{\mathrm{tgt}})$ of the target individual is predicted using the outputs $\boldsymbol{g}(\boldsymbol{I}^{\mathrm{tgt}})$ calculated as follows:

$$p(\boldsymbol{g}(\boldsymbol{I}^{\mathrm{tgt}}))|\boldsymbol{I}, \boldsymbol{Y}, \boldsymbol{I}^{\mathrm{tgt}}) \sim \mathcal{N}(\boldsymbol{\mu}^{\mathrm{tgt}}, \boldsymbol{\sigma}^{\mathrm{tgt}}), \tag{3.2.4}$$

$$\boldsymbol{\mu}^{\mathrm{tgt}} = \boldsymbol{K}^{\mathrm{tgt}\top}[\boldsymbol{K}(\boldsymbol{I}, \boldsymbol{I}) + \boldsymbol{\sigma}]^{-1}\boldsymbol{Y}, \tag{3.2.5}$$

$$\boldsymbol{\sigma}^{\mathrm{tgt}} = \boldsymbol{K}(\boldsymbol{I}^{\mathrm{tgt}}, \boldsymbol{I}^{\mathrm{tgt}}) - \boldsymbol{K}^{\mathrm{tgt}\top}[\boldsymbol{K}(\boldsymbol{I}, \boldsymbol{I}) + \boldsymbol{\sigma}]^{-1}\boldsymbol{K}^{\mathrm{tgt}}, \tag{3.2.6}$$

where $\boldsymbol{g}(\boldsymbol{I}^{\mathrm{tgt}}) = \mathrm{vec}(\boldsymbol{S}(p^{\mathrm{tgt}}, \boldsymbol{X}^{\mathrm{tgt}})^\top = [g_1(\boldsymbol{I}^{\mathrm{tgt}}), g_2(\boldsymbol{I}^{\mathrm{tgt}}), \ldots, g_T(\boldsymbol{I}^{\mathrm{tgt}})]^\top \in \mathbb{R}^T$, $\boldsymbol{K}^{\mathrm{tgt}} = \boldsymbol{K}(\boldsymbol{I}, \boldsymbol{I}^{\mathrm{tgt}}) \in \mathbb{R}^{CT \times T}$ consists of blocks $\boldsymbol{K}_{tt'}(\boldsymbol{I}, \boldsymbol{I}^{\mathrm{tgt}}) = [k_{tt'}(\boldsymbol{I}_c, \boldsymbol{I}^{\mathrm{tgt}})]_{c=1}^C \in \mathbb{R}^{C \times C}$, $\boldsymbol{\sigma} \in \mathbb{R}^{CT \times CT}$ is a diagonal noise matrix, and $\boldsymbol{K}(\boldsymbol{I}, \boldsymbol{I}) \in \mathbb{R}^{CT \times CT}$ and $\boldsymbol{K}(\boldsymbol{I}^{\mathrm{tgt}}, \boldsymbol{I}^{\mathrm{tgt}}) \in \mathbb{R}^{T \times T}$ include all $k_{tt'}(\boldsymbol{I}_c, \boldsymbol{I}_{c'})$. Under the independent assumptions, $a_q(\boldsymbol{I}_c) \perp a_{q'}(\boldsymbol{I}_c)$ and $b_t(\boldsymbol{I}_c) \perp b_{t'}(\boldsymbol{I}_c)$ with $q \neq q'$ and $t \neq t'$, the

covariance between $g_t(\boldsymbol{I}_c)$ and $g_{t'}(\boldsymbol{I}_{c'})$ is calculated as follows:

$$k_{tt'}(\boldsymbol{I}_c, \boldsymbol{I}_{c'}) = \begin{cases} \sum_{q=1}^{Q} w_{t,q}^2 k_q(\boldsymbol{I}_c, \boldsymbol{I}_{c'}) + k_t(\boldsymbol{I}_c, \boldsymbol{I}_{c'}) & t = t' \\ \sum_{q=1}^{Q} w_{t,q} w_{t',q} k_q(\boldsymbol{I}_c, \boldsymbol{I}_{c'}) & t \neq t', \end{cases} \tag{3.2.7}$$

where $k_q(\boldsymbol{I}_c, \boldsymbol{I}_{c'})$ represents the covariance between $a_q(\boldsymbol{I}_c)$ and $a_q(\boldsymbol{I}_{c'})$, and $k_t(\boldsymbol{I}_c, \boldsymbol{I}_{c'})$ is the covariance between $b_t(\boldsymbol{I}_c)$ and $b_t(\boldsymbol{I}_{c'})$. In this manner, CoMOGP considers the relationship among inter-outputs by calculating a_q and incorporates information on the output itself by calculating b_q. The elements of PSM are interconnected, and each element itself possesses unique features. These characteristics of CoMOGP make it well-suited for PSM prediction. In our proposed method, PSM can be predicted for each image based on the inputs of visual features into CoMOGP, and we can mitigate overfitting to a small amount of training data through the Gaussian process-based approach.

3.2.3 Experiments

3.2.3.1 Settings

This experiment adopted the same dataset and the same training strategy of the multi-task CNN as Section 3.1.3.1. From the dataset, 500 images were randomly chosen as test images, while the remaining 1100 images served as training images. Additionally, C (= 100) images were selected using AIS, representing those viewed by the target individuals. For this experiment, 10 participants were randomly designated as target individuals, leaving the remaining 20 as individuals with a substantial volume of gaze data. Besides, the USM, which was used in the multi-task CNN, was obtained as an average of the PSMs from individuals that were not designated as target individuals. Furthermore, for visual features, we employed the outputs of the final pooling layer of the Densenet201 model [73].

We conducted both quantitative and qualitative evaluations of our proposed method. To quantitatively assess the disparity between GT and the predicted PSM, we employed using CC, KL-div, and Sim presented in Section 2.2.3. In order to validate our proposed method, we compared

Table 3.2.1: Performance comparison across various evaluation indices. The symbol (↑) indicates that a higher index corresponds to improved performance, while the symbol (↓) indicates that a lower index reflects improved performance. It is important to mention that 100 (=C) selected images were utilized for training in PSM prediction methods. The use of bold font signifies the highest value within its respective evaluation index.

Methods	**Sim↑**	**KLdiv↓**	**CC↑**
Signature [23]	0.412	8.04	0.413
GBVS [22]	0.447	6.89	0.437
Itti [20]	0.391	9.04	0.322
SalGAN [29]	0.569	3.56	0.635
Baseline1 [72]	0.503	4.13	0.597
Baseline2 [71]	0.417	7.64	0.401
FPSP based on similarity [76]	0.401	1.82	0.735
FPSP using CoMOGP (Ours)	**0.655**	**1.38**	**0.765**

our method against four comparative methods, Signature [23], GBVS [22], Itti [20], and SalGAN [29], which are USM prediction methods selected from the MIT saliency benchmark [69]. It is worth noting that SalGAN uses deep learning with the SALICON dataset [70]. Additionally, we employed three PSM prediction methods designed for small amounts of gaze data.

Baseline1: PSM prediction utilizing relationships between parts of the image and the entire image [72]

Baseline2: PSM prediction relying on visual similarities [71]

PSM based on similarity: PSM prediction method presented in Chapter 3.1 [76]

It is important to mention that these PSM prediction methods were trained exclusively on the selected images via AIS.

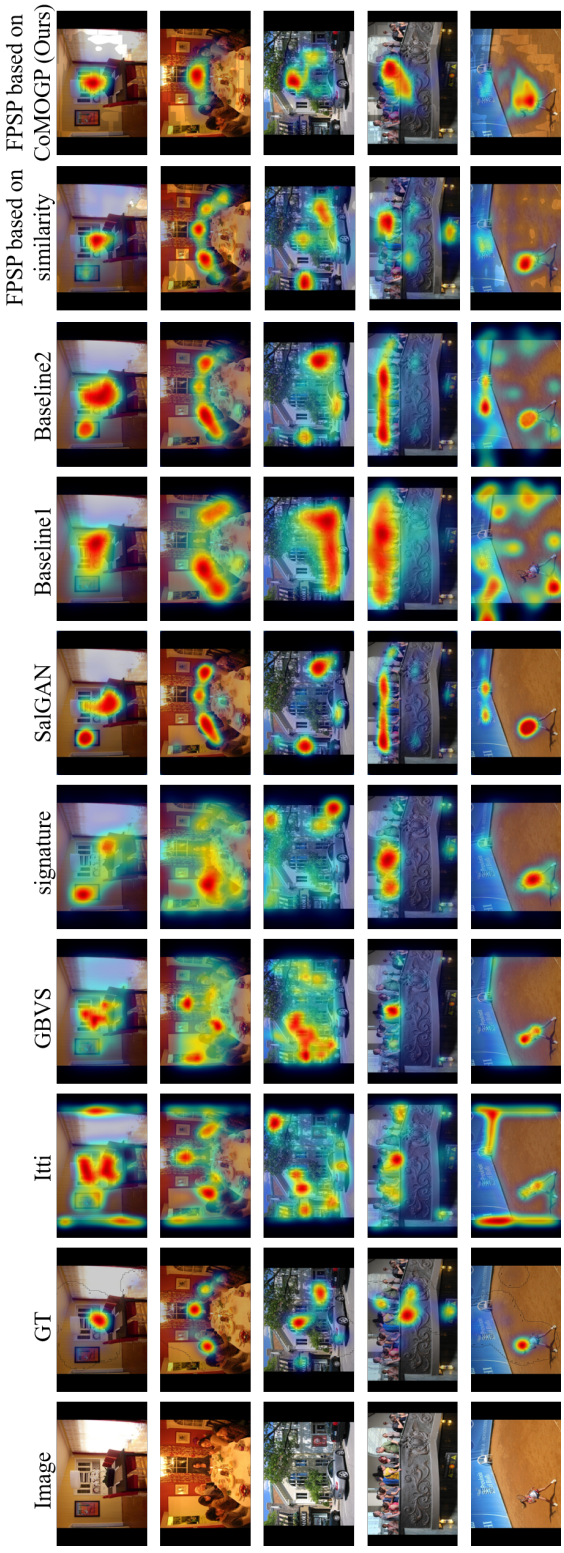

Figure 3.2.3: Qualitative outcomes for one individual predicted by the proposed and comparative methods.

3.2.3.2 Results and Discussion

Figure 3.2.3 and Table 3.2.1 present the quantitative and qualitative experimental results. In Fig. 3.2.3, the PSM that our method predicts exhibit the highest similarity to GT in comparison to the comparative methods. Table 3.2.1 further demonstrates the superiority of FPSP based on CoMOGP across all evaluation indices, outperforming all comparative methods. Specifically, when compared to Signature, GBVS, and Itti, our method showcases superiority over conventional USM prediction methods relying on computational models without training. Additionally, in comparison to SalGAN, our method underscores the efficacy of personalized prediction for PSM. Further comparison with Baselines1 and 2 highlights the effectiveness of leveraging relationships between the target individual and others, i.e., the use of PSMs predicted by the multi-task CNN is the effective approach for PSM prediction with a limited amount of training gaze data. Moreover, contrasting our approach with FPSP based on similarity reveals the effectiveness of adjusting weights based on the image and other individuals. In this way, the experimental results unequivocally validate the effectiveness of FPSP based on CoMOGP.

3.2.4 Conclusions

In this chapter, we present the FPSP approach employing individual similarity grounded in CoMOGP. Our proposed method accomplishes the prediction of PSM for the target individual with limited gaze data. The experimental findings demonstrate the efficacy of our approach through both quantitative and qualitative evaluations.

Chapter 3.3

Few-shot Personalized Saliency Prediction with Similarity of Gaze Tendency Using Object-based Structural Information

3.3.1 Introduction

In Chapter 3.2, we employ the collaborative multi-output Gaussian process regression (CoMOGP) [74] to predict the PSM with limited amount of training data, utilizing visual information from images and the predicted PSMs of training individuals. The CoMOGP-based approach incorporates the semantic information of images through the utilization of visual information. However, a drawback arises as this method transforms PSMs of training individuals into vectors for CoMOGP inputs, resulting in the loss of structural information within the PSMs. Given that visual saliency is significantly influenced by the structural characteristics of images [20], the absence of structural details from PSMs obtained through the multi-task CNN, which preserves image structures, imposes constraints on PSM prediction. Consequently, achieving highly accurate PSM prediction necessitates considering the semantic information of the image and retaining the structural information inherent in the PSMs.

In this chapter, we introduce a few-shot PSM prediction method incorporating gaze tendency similarities utilizing object-based structural information. In order to incorporate both semantic and structural information from images, our approach centers on the gaze tendency towards objects in images. Given the established connection between human gazes and objects [77], we

leverage similarities in gaze tendency for each object within images while comparing the target individual with the training individuals. As gaze data obtained for the target image is unavailable, we make the assumption that the target individual viewed a few images (hereafter referred to as common images) selected through the AIS scheme. Our proposed method then seeks visually similar objects among those present in the common images. Specifically, we conduct object detection in both common and target images. Subsequently, gaze data corresponding to common images is collected from both training and target individuals, allowing us to calculate the gaze tendency similarities for objects present in common images. Following this, visual similarities of objects between common and target images are computed to seek the similar objects. Consequently, the PSMs for each object in the target image are predicted based on the PSMs predicted for the training individuals and the gaze tendency similarities for analogous objects. Through these processes, gaze tendency similarities are computed for each image utilizing object-based visual similarities, enabling our method to incorporate semantic information. Furthermore, structural information is taken into account by integrating the PSMs predicted for training individuals while preserving their structures. This chapter offers a distinctive contribution by directing attention to the gaze tendency similarities for visually akin objects. This emphasis aims to enhance the efficacy of the PSM prediction method, particularly when working with a limited set of gaze data, by concurrently incorporating both semantic and structural information.

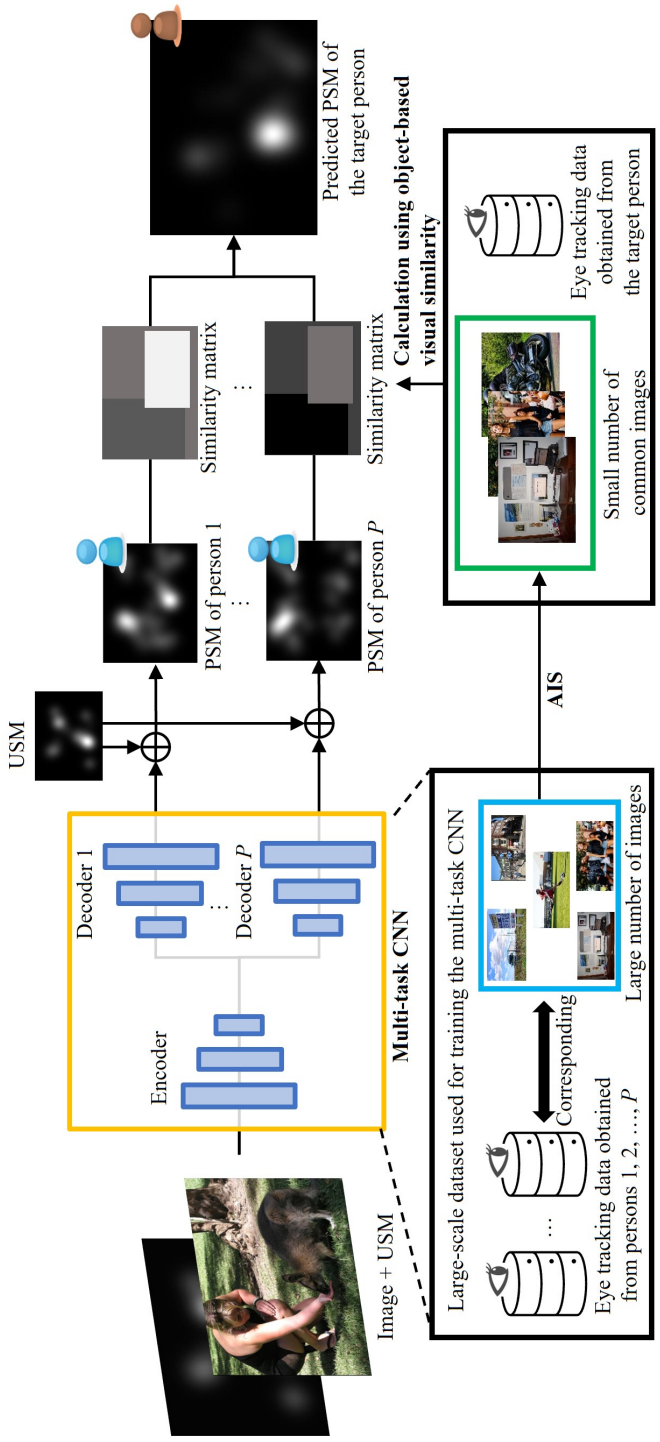

Figure 3.3.1: Flow of the proposed method. Initially, a multi-task CNN is employed to predict the PSMs for the training individuals, denoted as $1, 2, \ldots, P$. Subsequently, utilizing the AIS scheme, we identify common images that the target individual should view. Lastly, we amalgamate the predicted PSMs by incorporating the gaze tendency similarity between the target individual and the training individuals, leveraging object-based visual similarity.

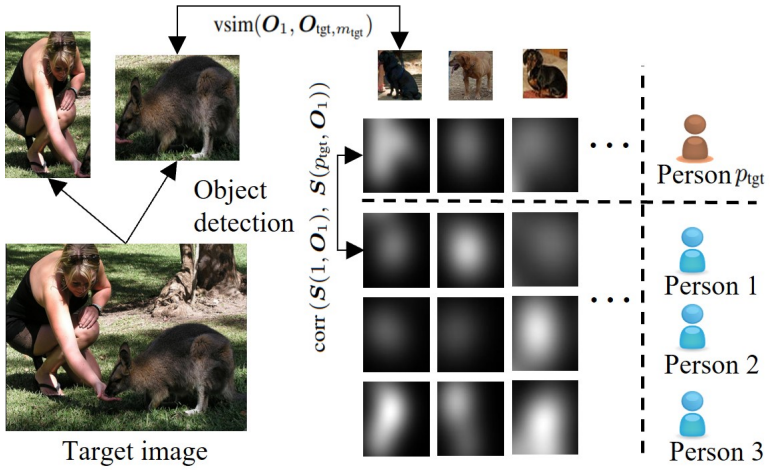

Figure 3.3.2: Computation way of object-based visual similarity. Starting with the target image, we initially identify objects and subsequently retrieve analogous objects from those present in the common images. Furthermore, the calculation of gaze tendency similarities among the target and training individuals relies on the PSM for the retrieved objects.

3.3.2 Proposed Few-shot PSM Prediction

This section describes the intricacies of the proposed few-shot PSM prediction method comprising three fundamental steps as illustrated in Fig. 3.3.1. Initially, we train a multi-task CNN to predict PSMs for individuals with substantial training data. Subsequently, employing the AIS scheme, we selectively identify common images known for inducing more varied gazing patterns than other images. Finally, the few-shot PSM prediction is executed by leveraging gaze tendency similarities based on object-based visual similarity. In this section, calculation of object-based gaze similarity is mainly explained, while multi-task CNN and AIS are explained in Sections 3.1.2.1 and 3.1.2.2, respectively.

3.3.2.1 PSM Prediction with Object-based Gaze Similarity

In this section, we delineate the computation of gaze tendency similarities between the target individual and training individuals using object-based visual similarity, along with the method for predicting the PSM of the target individual. The calculation of gaze tendency similarities relies on the set of common images $\boldsymbol{X}_c$ ($c = 1, 2, \ldots, C$; C representing the number of selected images) chosen by AIS. It is crucial to note that we presume gaze data for the common images

are available for both the target individual and training individuals, allowing the calculation of PSMs $\boldsymbol{S}(p, \boldsymbol{X}_c)$ and $\boldsymbol{S}(p_{\mathrm{tgt}}, \boldsymbol{X}_c)$ for individual p and the target individual, respectively. In this context, we assess the similarities in gaze tendency for the entire image and the objects within the target image. Specifically, we compute the gaze tendency similarity matrix $\boldsymbol{W}_{\mathrm{all},p}$ between individual p and the target individual for the overall images as follows:

$$\boldsymbol{W}_{\mathrm{all},p} = \frac{\sum_{c=1}^{C} \mathrm{corr}\left(\boldsymbol{S}(p, \boldsymbol{X}_c),\ \boldsymbol{S}(p_{\mathrm{tgt}}, \boldsymbol{X}_c)\right)}{\sum_{p'=1}^{P}\sum_{c=1}^{C} \mathrm{corr}\left(\boldsymbol{S}(p', \boldsymbol{X}_c),\ \boldsymbol{S}(p_{\mathrm{tgt}}, \boldsymbol{X}_c)\right)}, \tag{3.3.1}$$

where $\mathrm{corr}\,(\cdot,\ \cdot)$ denotes the Pearson's correlation coefficient. Subsequently, for computing the gaze tendency similarity concerning the objects within the target image, we identify the objects $\boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}}$ ($m_{\mathrm{tgt}} = 1, 2, \ldots, M_{\mathrm{tgt}}$; M_{tgt} indicating the number of objects in the target image) using the object detection method [78]. Determining the gaze tendency similarity for objects in unseen images, such as the target image in this context, poses a challenge. To overcome this, we apply the similarity of the PSMs for analogous objects in the common images to that of the target image, as illustrated in Fig. 3.3.2. Subsequently, we identify objects $\boldsymbol{O}_{c,m_c}$ ($m_c = 1, 2, \ldots, M_c$; M_c representing the number of objects in the cth common image) in the common images $\boldsymbol{X}_c$ using the same approach as for the target image. To determine similar objects, we assess the visual similarity between objects $\boldsymbol{O}_{c,m_c}$ and $\boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}}$ employing the function f_v for extracting visual features, defined as follows:

$$\mathrm{vsim}(\boldsymbol{O}_{c,m_c}, \boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}}) = \mathrm{dis}(f_v(\boldsymbol{O}_{c,m_c}), f_v(\boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}})), \tag{3.3.2}$$

where $\mathrm{dis}(\cdot,\cdot)$ represents the distance between visual features. Additionally, by utilizing the PSMs for the J most similar objects, the gaze tendency similarities for the object $\boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}}$ are calculated as follows:

$$\boldsymbol{W}_{m_{\mathrm{tgt}},p} = \frac{\sum_{j=1}^{J} \mathrm{vsim}(\boldsymbol{O}_j, \boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}})\tilde{\boldsymbol{W}}_{m_{\mathrm{tgt}},p,j}}{\sum_{p'=1}^{P}\sum_{j=1}^{J} \mathrm{vsim}(\boldsymbol{O}_j, \boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}})\tilde{\boldsymbol{W}}_{m_{\mathrm{tgt}},p',j}}, \tag{3.3.3}$$

$$\tilde{\boldsymbol{W}}_{m_{\mathrm{tgt}},p,j} = \mathrm{corr}\left(\boldsymbol{S}(p, \boldsymbol{O}_j),\ \boldsymbol{S}(p_{\mathrm{tgt}}, \boldsymbol{O}_j)\right), \tag{3.3.4}$$

where $\boldsymbol{O}_j$ is the jth most similar object to the object $\boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}}$. Finally, we calculate the gaze tendency similarity matrix $\boldsymbol{W}_p$ by applying $\boldsymbol{W}_{m_{\mathrm{tgt}},p}$ and $\boldsymbol{W}_{\mathrm{all},p}$ to the region of the object $\boldsymbol{O}_{\mathrm{tgt},m_{\mathrm{tgt}}}$

and other regions such as backgrounds, respectively. It is noteworthy that we apply the mean of $\boldsymbol{W}_{m_{\mathrm{tgt}},p}$ to the regions where objects overlap.

We integrate PSMs $\hat{\boldsymbol{S}}(p, \boldsymbol{X}_{\mathrm{tgt}})$ predicted by the multi-task CNN for predicting the PSM of the target individual, utilizing the gaze tendency similarity matrix as follows:

$$\hat{\boldsymbol{S}}(p_{\mathrm{tgt}}, \boldsymbol{X}_{\mathrm{tgt}}) = \sum_{p=1}^{P} \hat{\boldsymbol{S}}(p, \boldsymbol{X}_{\mathrm{tgt}}) \odot \boldsymbol{W}_{p}. \tag{3.3.5}$$

Therefore, the proposed method enables the simultaneous consideration of both semantic and structural information in Eqs. (3.3.3) and (3.3.5), realizing few-shot PSM prediction with high accuracy using the object information in gaze tendency similarities.

3.3.3 Experiments

3.3.3.1 Settings

This experiment adopted the same dataset and the same training strategy of the multi-task CNN as Section 3.1.3.1. Subsequently, we randomly partitioned the dataset into 500 test images and 1100 training images, selecting C (= 100) common images from the training images using the AIS scheme. Notably, individuals in the PSM dataset were randomly divided into 10 target individuals and 20 training individuals. Target individuals exclusively viewed common images, while training individuals viewed the training images. Additionally, the USM, which was used in the multi-task CNN, was calculated as the average of the PSMs of the training individuals to mitigate the impact of USM calculation errors. Visual features and the distance metric $\mathrm{dis}(\cdot, \cdot)$ utilized the outputs of the final pooling layer of the Inception-Resnet-v2 model [79] and the standardized Euclidean distance, respectively. Moreover, we set $J = 5$.

To assess the effectiveness of our method, both quantitative and qualitative evaluations were performed. For quantitative evaluation against GTs, using CC, KLdiv, and Sim presented in Section 2.2.3 were employed as evaluation metrics [35]. In this experiment, we compared our method with the following existing methods, Signature [23], GBVS [22], Itti [20], SalGAN [29], and Contextual [80], which are USM prediction methods from the MIT saliency benchmark [69]. SalGAN and Contextual were trained with the SALICON dataset [70]. Additionally, we consid-

Table 3.3.1: Performance comparison across various evaluation indices. The symbol (↑) indicates that a higher index corresponds to improved performance, while the symbol (↓) indicates that a lower index reflects improved performance. It is important to mention that 100 (=C) selected images were utilized for training in PSM prediction methods. The use of bold font signifies the highest value within its respective evaluation index.

Methods	**Sim↑**	**KLdiv↓**	**CC↑**
Signature [23]	0.412	8.04	0.413
GBVS [22]	0.447	6.89	0.437
Itti [20]	0.391	9.04	0.322
SalGAN [29]	0.569	3.56	0.635
Contextual [80]	0.580	3.57	0.674
Baseline1 [72]	0.503	4.13	0.597
Baseline2 [71]	0.417	7.64	0.401
Similarity-based FPSP [76]	0.401	1.82	0.735
CoMOGP-based FPSP [81]	**0.655**	1.38	0.765
Proposed Method	0.642	**1.09**	**0.781**

ered four PSM prediction methods using a small amount of gaze data:

Baseline1: PSM prediction using local and global information of input images [72].

Baseline2: PSM prediction using visual similarities of the target and training images [71].

Similarity-based FPSP: PSM prediction method presented in Chapter 3.1 [76].

CoMOGP-based FPSP: PSM prediction method presented in Chapter 3.2 [81].

It is important to note that all PSM prediction methods were trained exclusively with the common images selected by AIS.

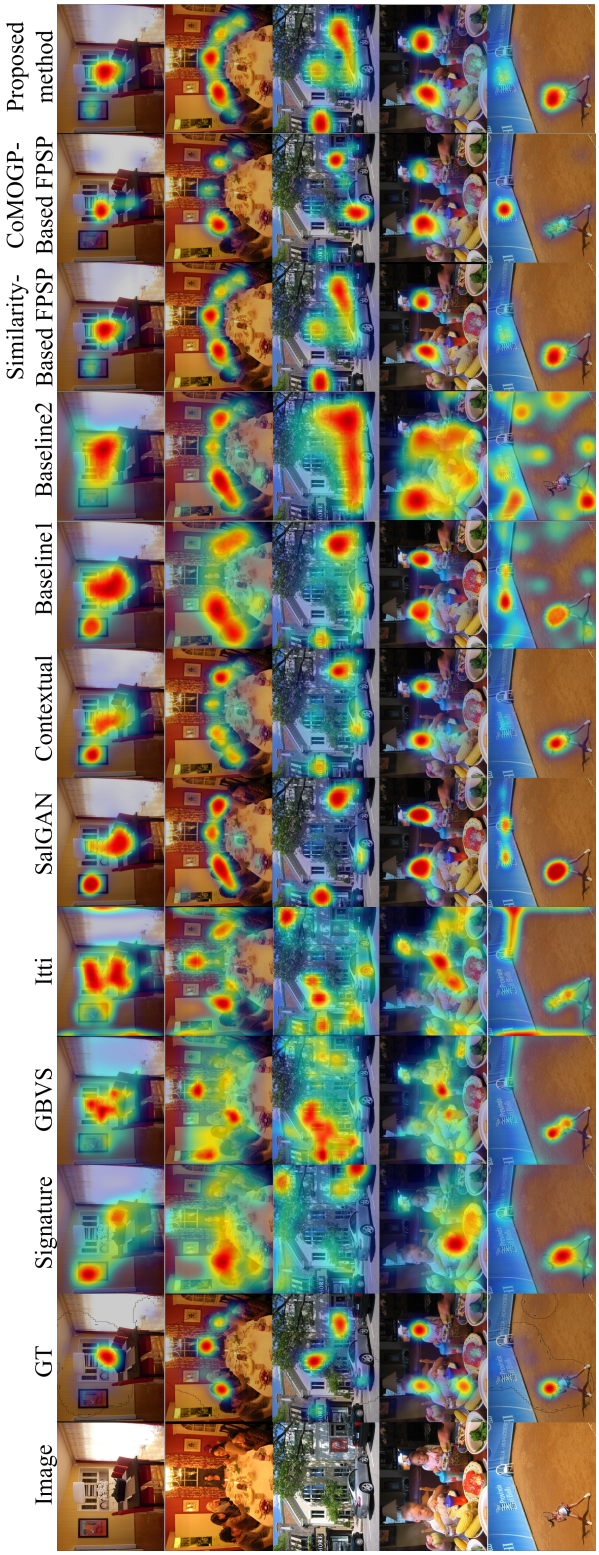

Figure 3.3.3: Qualitative outcomes for one individual predicted by the proposed and comparative methods.

3.3.3.2 Performance Evaluation

We present the experimental results in Fig. 3.3.3 and Table 3.3.1. Specifically, Fig. 3.3.3 visually illustrates that the PSM predicted by our proposed method closely aligns with the GT, showcasing the qualitative efficacy of our approach. For a quantitative assessment, we compare our method with others in Table 3.3.1. Our method surpasses all comparative methods across all metrics, exhibiting superior performance to CoMOGP-based FPSP [81] in "CC" and "KL-div". This validates the effectiveness of our proposed method. More precisely, a comparison with the state-of-the-art USM prediction method, Contextual [80], underscores the efficacy of personalized prediction. Furthermore, a comparison with Similarity-based FPSP [76] highlights the efficiency of incorporating the gaze tendency similarity based on object information. Next, a comparison with CoMOGP-based FPSP [81], which can consider semantic information, reveals that our method outperforms it in most evaluation metrics, emphasizing the efficacy of leveraging structural information. In the end, by comparing the proposed FPSP with both CoMOGP-based FPSP [81] and Similarity-based FPSP [76], we emphasize the efficiency of simultaneously incorporating structural and semantic information.

3.3.4 Conclusions

This chapter introduces a method for predicting PSMs in a few-shot scenario by leveraging gaze tendency similarity through object-based structural information. The emphasis of the proposed approach lies in aggregating PSMs predicted for different individuals, addressing the scarcity of gaze data. Through experiments on an open dataset, the proposed method demonstrates superior performance compared to other approaches

Chapter 4

Gaze-based Emotional Category Classification

Given the widespread availability of web images, there is a growing need for comprehensive image analysis [82, 83]. Understanding of image contents primarily involves two key aspects: image-based information and human-based information. Many studies have leveraged image-based information to accomplish tasks such as semantic segmentation and object recognition [4, 84–87]. Simultaneously, human-based information has been employed for interest level estimation and image emotion recognition [19, 43]. Therefore, we distinguish understanding of image contents into two main categories: image-based understanding and human-based understanding, aligning with the first and second types of information, respectively. Despite the advancements facilitated by Convolutional Neural Networks (CNNs) [4] in achieving high-performance image-based understanding [4, 84–87], human-based understanding remains challenging, given its intricate connection to abstract semantics perceived by humans [88]. Specifically, image emotions represent the highest level of abstract semantics, defined as descriptors capturing the types and intensities of feelings, sensibility, moods, or affections experienced by humans when viewing images [89]. Therefore, this chapter centers on the image classification into emotional categories. In research on estimating emotions when humans view images, the efficiency of utilizing various types of biological data has been validated [16–18]. In the fields of psychological and neuroscience, it has demonstrated that objects present in images relate to human emotions [90, 91]. Additionally, a correlation exists between the emotional attributes of

images and the temporal changes in visual attention, which are intricately linked to human emotions [77]. Therefore, akin to emotion estimation, incorporating information about the viewed objects and temporal changes in visual attention is expected to be effective for the image classification of emotional categories.

Preliminaries

In this chapter, we employ specific mathematical notations to elucidate tensor analysis. The tensor order aligns with the count of modes. Throughout this chapter, individual lowercase letters such as a denote scalars, boldface lowercase letters such as $\boldsymbol{a}$ denote vectors (first-order tensors), boldface capital letters such as $\boldsymbol{A}$ denote matrices (second-order tensors), and calligraphic letters such as $\mathcal{X}$ denote tensors (third-order tensors or higher tensors).

The mode-l matricization of a kth-order tensor $\mathcal{X}^{k\text{th}} \in \mathbb{R}^{D_1 \times D_2 \times \cdots D_k}$ is denoted by $\text{mat}_l(\mathcal{X}^{k\text{th}}) \in \mathbb{R}^{D_l \times \prod i \neq l D_i}$. This is an ensemble of vectors in $\mathbb{R}^{m_l}$ obtained by holding the lth mode fixed and varying the other modes. The mode-l product of a kth-order tensor $\mathcal{X}^{k\text{th}}$ is denoted as $\mathcal{X}^{k\text{th}} \times_l \boldsymbol{Y}_l \in \mathbb{R}^{D_1 \times D_2 \times \cdots \times D_{l-1} \times D_l^* \times D_{l+1} \times \cdots \times D_k}$ using a matrix $\boldsymbol{Y}_l \in \mathbb{R}^{D_l \times D_l^*}$. Multiple multiplications are succinctly expressed as follows:

$$\mathcal{X}^{k\text{th}} \times_{\bar{l}} \boldsymbol{Y}_l = \mathcal{X}^{k\text{th}} \times_1 \boldsymbol{Y}_1 \times_2 \boldsymbol{Y}_2 \times \cdots \times_{l-1} \boldsymbol{Y}_{l-1} \times_{l+1} \boldsymbol{Y}_{l+1} \times \cdots \times_k \boldsymbol{Y}_k. \tag{4.1}$$

Moreover, the expression $\langle \mathcal{X}, \mathcal{Y} \rangle$, where the size of $\mathcal{Y}$ matches that of $\mathcal{X}$, denotes the inner product. These notations align with those utilized in prior studies [92, 93].

Chapter 4.1

Estimation of Emotion Labels via Tensor-based Spatiotemporal Visual Attention Analysis

4.1.1 Introduction

We introduce an approach for estimating emotion labels by employing tensor-based analysis for spatiotemporal visual attention utilizing gaze data in this chapter. Our approach involves the creation of a fourth-order Gaze and Image Tensor (GIT) that incorporates the target image and gaze data, establishing associations between images and the temporal evolution of visual attention, as depicted in Fig. 4.1.1. The first and second modes correspond to pixel locations, the third mode represents color channels, and the fourth mode captures changes in visual attention over time. Subsequently, we develop two neural networks designed for estimating emotion labels by considering temporal changes in visual attention and the objects included in target image, as depicted in Fig. 4.1.2. The first network directly leverages the fourth-order GIT, allowing us to incorporate spatial structures of visual attention across temporal changes. To achieve this, we employ supervised feature transformation through General Tensor Discriminant Analysis (GTDA) [92] on the fourth-order GIT. This calculates highly discriminative features for estimating emotion labels and classifies the features, which are calculated by GTDA, using Extreme Learning Machine (ELM) [94], enabling efficient training with a limited number of training samples. The second network partitions the tensor at each timestep and engage in transfer learning utilizing features obtained from a pre-trained Convolutional Neural Network [4] (CNN

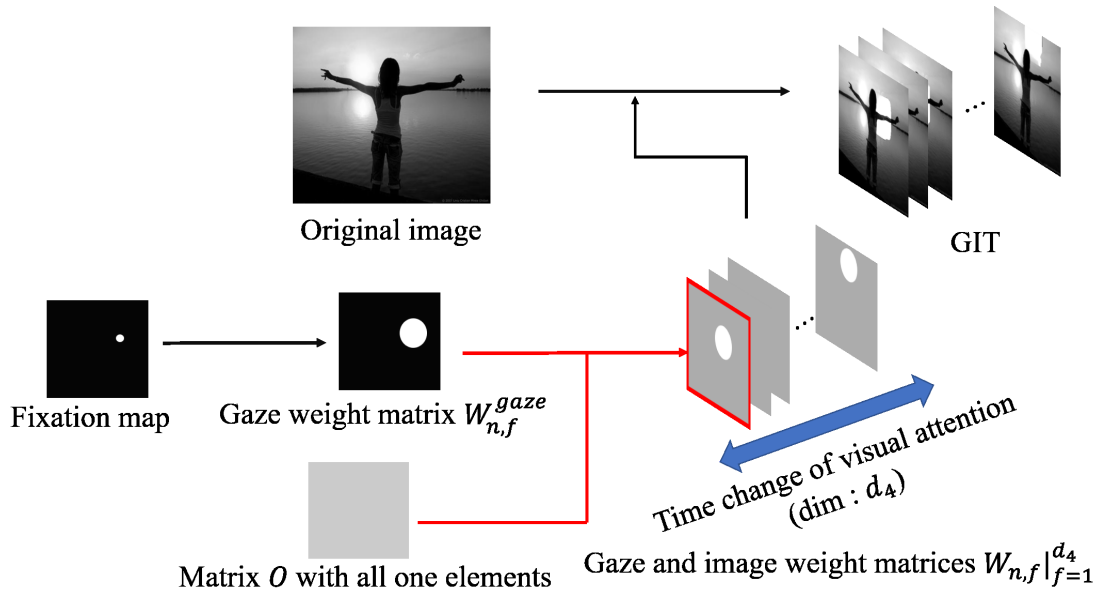

Figure 4.1.1: Overview of the GIT construction. While our approach deals with color images and builds a fourth-order GIT, this illustration depicts a gray-scale image for visual simplicity.

features), recognized for its significant contribution to object recognition [95]. To specifically capture visual features from objects pertinent to human emotions, CNN features are extracted from each frame of the fourth-order GIT. Through the alignment of these CNN features, we derive the second-order GIT, subjecting it to GTDA and constructing an ELM following the same procedure as the first network. At the end, the proposed method conducts emotion label estimation through the decision-level fusion of outputs from both networks. This approach facilitates emotion label estimation through tensor-based spatiotemporal visual attention analysis.

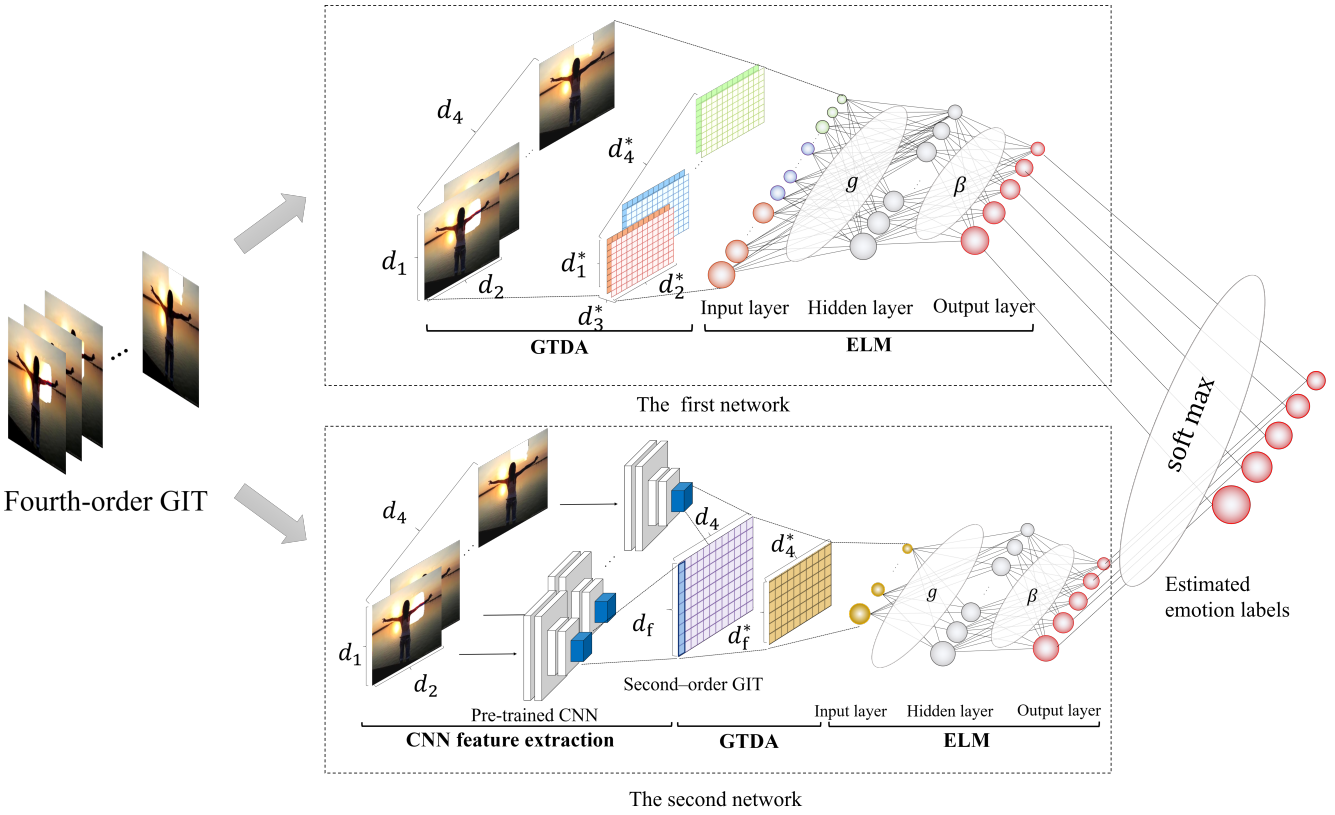

Figure 4.1.2: Overview of our approach. Initially, we formulate the fourth-order GIT to serve as the input for our proposed network. The network consists of two components. The first one addresses the temporal evolution of visual attention, while the second one emphasizes object characteristics. Subsequently, our method conducts emotion label estimation by strategically fusing the outputs from these networks

4.1.2 Our Emotion Label Estimation

The proposed method involves estimating emotion labels through tensor-based spatiotemporal visual attention analysis. Initially, the method introduces a fourth-order GIT, as illustrated in Fig. 4.1.1, and employs two neural networks depicted in Fig. 4.1.2 for emotion label estimation.

4.1.2.1 GIT Construction

In our proposed approach, we form the fourth-order GIT using gaze data, which includes both gaze coordinates and their corresponding duration times. For given training images $\mathcal{X}_n^{\text{img}} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ ($n = 1, 2, \ldots, N$; N denotes the number of training images, d_1 and d_2 represent the width and height of an image, respectively, and d_3 indicates the number of color channels), we generate a fixation map for each frame based on gaze data. Subsequently, we apply a Gaussian filter to the fixation map of each frame f ($= 1, 2, \ldots, d_4$; d_4 being the number of frames). The total gaze duration time for one image is divided into d_4 segments. We then compute a gaze weight matrix $\boldsymbol{W}_{n,f}^{\text{gaze}} \in \mathbb{R}^{d_1 \times d_2}$ corresponding to each frame f of the GIT. Additionally, we derive a gaze and image weight matrix $\boldsymbol{W}_{n,f}$ using the following formula:

$$\boldsymbol{W}_{n,f} = \boldsymbol{O} + d_4 \frac{\boldsymbol{W}_{n,f}^{\text{gaze}}}{\sum_{f=1}^{d_4} \boldsymbol{W}_{n,f}^{\text{gaze}}}, \tag{4.1.1}$$

where $\boldsymbol{O} \in \mathbb{R}^{d_1 \times d_2}$ is a matrix with all elements set to one. Consequently, we compute the fourth-order GIT $\mathcal{X}_n^{\text{4th}} \in \mathbb{R}^{d_1 \times d_2 \times d_3 \times d_4}$ based on gaze data as follows:

$$\boldsymbol{X}_{n,col,f}^{\text{4th}} = \boldsymbol{X}_{n,col}^{\text{img}} \circ \boldsymbol{W}_{n,f}, \tag{4.1.2}$$

where $\boldsymbol{X}_{n,col}^{\text{img}} \in \mathbb{R}^{d_1 \times d_2}$ ($col = 1, 2, \ldots, d_3$) is a segment of $\mathcal{X}_n^{\text{img}}$, and $\boldsymbol{X}_{n,col,f}^{\text{4th}} \in \mathbb{R}^{d_1 \times d_2}$ is a segment of $\mathcal{X}_n^{\text{4th}}$. Note that "$\circ$" denotes the Hadamard product operator. For gray-scale images, we treat the same values for each channel, constructing the fourth-order GIT. Consequently, we achieve the construction of GIT, which effectively represents images with consideration for changes in visual attention.

4.1.2.2 CNN Feature Extraction

To calculate visual features from viewed objects, our proposed method leverages the outputs of the final pooling layer in a pre-trained CNN. Typically, the images fed into the CNN possess three dimensions corresponding to pixel location coordinates and color channels, and our method calculates CNN features from each frame f to construct the second-order GIT $\boldsymbol{X}_n^{\mathrm{2nd}} \in \mathbb{R}^{d_f \times d_4}$ (where d_f represents the dimension of CNN features) by aligning the computed visual features. Given that CNN features play a pivotal role in object recognition, the resultant second-order GIT becomes closely linked with objects. Thus, in the second network, we employ CNN features to discern the characteristics of focused objects in images. The procedures outlined in this section correspond to the "CNN feature extraction" depicted in Fig. 4.1.2.

4.1.2.3 GTDA-based Feature Transformation

Given the common application of GTDA to both the fourth-order GIT and the second-order GIT, which are derived in the upper and lower networks in Fig. 4.1.2, we introduce a kth-order tensor $\mathcal{A}_{i;j}^{k\mathrm{th}} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_k}$ representing these two types of GITs. Hence, in this section, we replace $\boldsymbol{X}_n^{\mathrm{2nd}}$ and $\mathcal{X}_n^{\mathrm{4th}}$ with $\mathcal{A}_{i;j}^{k\mathrm{th}}$ for simplicity. It is noteworthy that $\mathcal{A}_{i;j}^{k\mathrm{th}}$ is the jth training tensor in the ith individual class ($j = 1, 2, \ldots, n_i$; n_i being the number of training tensors in the ith class, $i = 1, 2, \ldots, c$; c being the number of classes). To acquire the transformation set $\{\boldsymbol{P}_l^* \in \mathbb{R}^{m_l \times m_l^*}\}_{l=1}^k$ ($m_l^* < m_l$), GTDA addresses the optimization problem:

$$\{\boldsymbol{P}_l^*\}_{l=1}^k = \underset{\{\boldsymbol{P}_l\}_{l=1}^k}{\arg\max}\, \mathrm{tr}\left(\boldsymbol{P}_l^\top \left(\boldsymbol{C}_l^b - \zeta_l \boldsymbol{C}_l^w\right) \boldsymbol{P}_l\right), \tag{4.1.3}$$

where ζ_l serves as a tuning parameter. In our proposed method, ζ_l equals the maximum eigenvalue of $(\boldsymbol{C}_l^w)^{-1}\boldsymbol{C}_l^b$ as described in [92]. Furthermore, we define $\boldsymbol{C}_l^b$ and $\boldsymbol{C}_l^w$ as follows:

$$\boldsymbol{C}_l^b = \sum_{i=1}^{c} \left[n_i \mathrm{mat}_l\left((\mathcal{M}_i - \mathcal{M}) \times_{\bar{l}} \boldsymbol{P}_l^\top\right) \mathrm{mat}_l^\top\left((\mathcal{M}_i - \mathcal{M}) \times_{\bar{l}} \boldsymbol{P}_l^\top\right)\right], \tag{4.1.4}$$

$$\boldsymbol{C}_l^w = \sum_{i=1}^{c} \sum_{j=1}^{n_i} \left[\mathrm{mat}_l\left(\left(\mathcal{A}_{i;j}^{k\mathrm{th}} - \mathcal{M}_i\right) \times_{\bar{l}} \boldsymbol{P}_l^\top\right) \mathrm{mat}_l^\top\left(\left(\mathcal{A}_{i;j}^{k\mathrm{th}} - \mathcal{M}_i\right) \times_{\bar{l}} \boldsymbol{P}_l^\top\right)\right], \tag{4.1.5}$$

where $\mathcal{M}_i = (1/n_i)\sum_{j=1}^{n_i}\mathcal{A}_{i;j}^{k\text{th}}$ denotes the class mean tensor for the ith class. Additionally, $\mathcal{M} = (1/N)\sum_{i=1}^{C} n_i\mathcal{M}_i$ stands for the total mean tensor, aggregating all training tensors. It is important to note that $\mathcal{A}_{i;j}^{k\text{th}} \mid_{1\le i\le c}^{1\le j\le n_i}$, $\mathcal{M}_i \mid_{i=1}^{c}$, and $\mathcal{M}$ are all kth-order tensors residing in $\mathbb{R}^{m_1\times m_2\times\cdots\times m_k}$. Finally, a tensor $\mathcal{B}_{i;j}^{k\text{th}}$ is derived by transforming the kth-order tensor $\mathcal{A}_{i;j}^{k\text{th}}$ according to the following equation:

$$\mathcal{B}_{i;j}^{k\text{th}} = \mathcal{A}_{i;j}^{k\text{th}} \prod_{l=1}^{k} \times_l \boldsymbol{P}_l^*. \tag{4.1.6}$$

This process enables the computation of highly discriminative features for emotion label estimation through the application of GTDA, taking into account the label information.

4.1.2.4 ELM-based Emotion Label Estimation

This section describes the way to construct classifiers based on ELM and the emotion label estimation grounded on the dual outputs of the proposed networks. The ELM-based classifiers undergo training utilizing the transformed tensor $\boldsymbol{Y}_n^{2\text{nd}} \in \mathbb{R}^{d_f^*\times d_4^*}$ and $\mathcal{Y}_n^{4\text{th}} \in \mathbb{R}^{d_1^*\times d_2^*\times d_3^*\times d_4^*}$. Tensors $\boldsymbol{Y}_n^{2\text{nd}}$ and $\mathcal{Y}_n^{4\text{th}}$ result from the application of GTDA to $\boldsymbol{X}_n^{2\text{nd}}$ and $\mathcal{X}_n^{4\text{th}}$, respectively, analogous to the role of $\mathcal{B}_{i;j}^{k\text{th}}$ in the preceding section. It is pertinent to note that d^* represents the number of transformed dimensions through GTDA, corresponding to m^* in the preceding section. ELM comprises a three-layer neural network and trains a weight matrix connecting a hidden layer and an output layer through the subsequent equation:

$$\boldsymbol{\beta} = \boldsymbol{Z}^{\dagger}\boldsymbol{T}, \tag{4.1.7}$$

where $\boldsymbol{T} = [\boldsymbol{t}_1^\top, \boldsymbol{t}_2^\top, \ldots, \boldsymbol{t}_n^\top]^\top$. It is noteworthy that $\boldsymbol{t}_n = [t_{n,1}, t_{n,2}, \ldots, t_{n,c}]$ is a one-hot encoded class vector. By using the one-hot encoding, the element corresponding to the class of the nth training image is one, while the rest of the elements are zeros. Besides, $\boldsymbol{Z}^{\dagger}$ is the Moore-Penrose generalized inverse [96] of the output matrix $\boldsymbol{Z}$. The output matrix of a hidden layer of ELM is calculated as follows:

$$\boldsymbol{Z} = [z(\boldsymbol{y}_1), z(\boldsymbol{y}_2), \ldots, z(\boldsymbol{y}_N)]^\top, \tag{4.1.8}$$

where, $\boldsymbol{y}_n$ is the vectorization of the transformed tensors $\boldsymbol{Y}_n^{\text{2nd}}$ or $\mathcal{Y}_n^{\text{4th}}$. $\boldsymbol{z}(\boldsymbol{y}_n)$ is obtained by applying an activation function g to $\boldsymbol{y}_n$ as follows:

$$\boldsymbol{z}(\boldsymbol{y}_n) = [g(\boldsymbol{a}_1^\top \boldsymbol{y}_n + b_1), g(\boldsymbol{a}_2^\top \boldsymbol{y}_n + b_2), \ldots, g(\boldsymbol{a}_E^\top \boldsymbol{y}_n + b_E)]^\top. \tag{4.1.9}$$

Notably, E represents the number of neurons in the hidden layer. Additionally, $\boldsymbol{a}_e$ and b_e ($e = 1, 2, \ldots, E$) serve as parameters for the activation function g, with $\boldsymbol{a}_e$ and b_e being random values obtained from a uniform distribution. Given a test vector $\boldsymbol{y}$ obtained from a transformed test tensor $\boldsymbol{Y}_n^{\text{2nd}}$ or $\mathcal{Y}_n^{\text{4th}}$, the output of ELM is calculated as follows:

$$f(\boldsymbol{y}) = \boldsymbol{z}(\boldsymbol{y})^\top \boldsymbol{\beta}. \tag{4.1.10}$$

In our proposed approach, we construct the initial classifier utilizing the transformed fourth-order GIT $\mathcal{Y}_n^{\text{4th}}$ as input, and the secondary classifier utilizing the transformed second-order GIT $\boldsymbol{Y}_n^{\text{2nd}}$ as input. The initial network predicts emotion labels by considering the temporal progression of visual attention, whereas the secondary network predicts emotion labels by paying attention to specific objects in the images. In the end, we employ a softmax function for decision-level fusion, combining the results of the two classifiers, and determining the emotion label based on the outputs of the softmax function.

4.1.3 Experiments

This section presents the experimental results validating the effectiveness of our proposed method. The experiments utilized the abstract paintings dataset [39], comprising 280 images, and each image is labeled with at least one emotion from eight categories (*amusement, awe, contentment, excitement, anger, disgust, fear*, and *sad*). The ground truth (GT) for each image was derived from emotion labels assigned by around 14 individuals. Our method was applied to estimate each emotion label, using 224 randomly selected images for training and remaining 56 images for testing. The performance evaluation employed F-measure, the harmonic mean of Recall and Precision. To carefully consider the data imbalance, the number of images for each class was equalized through random selection and the proposed method was trained for

each emotion label. Tobii Eye Tracker 4C[1] was used in this experiment with 13 participating participants (Pars 1-13). Participants were tasked with viewing each image until recalling some emotions, with one second allocated to adjust their gaze to the center of the monitor before viewing each image. The length of gazing time was normalized as it became d_4.

For the determination of the number of hidden neurons in ELM, a four-fold cross-validation was conducted on the training dataset. The optimal number of hidden neurons, providing the best estimation performance, was selected based on the validation dataset. A sigmoid function served as the activation function g of ELM.

The following seven comparative methods (CMs) were adopted for comparing them with the proposed method (PM):

- CM1

  This method ignores the temporal evolution of visual attention, constructing the fourth-order GIT of $d_4 = 1$ (i.e., third order GIT) and extracting CNN features directly from the third order GIT.

- CM2

  Similar to PM, this method has two networks. The first network extracts gaze features, inputting them directly into ELM. The gaze features align with a gaze analysis method [97]. The second network employs the approach from CM1.

- CM3

  This method utilizes only the first network of PM.

- CM4

  This method utilizes only the second network of PM.

- CM5

  A baseline method [47] using hand-crafted visual features, which are calculated by applying Gabor and Sobel filters, and gaze information .

- CM6

[1] https://tobiigaming.com/eye-tracker-4c/

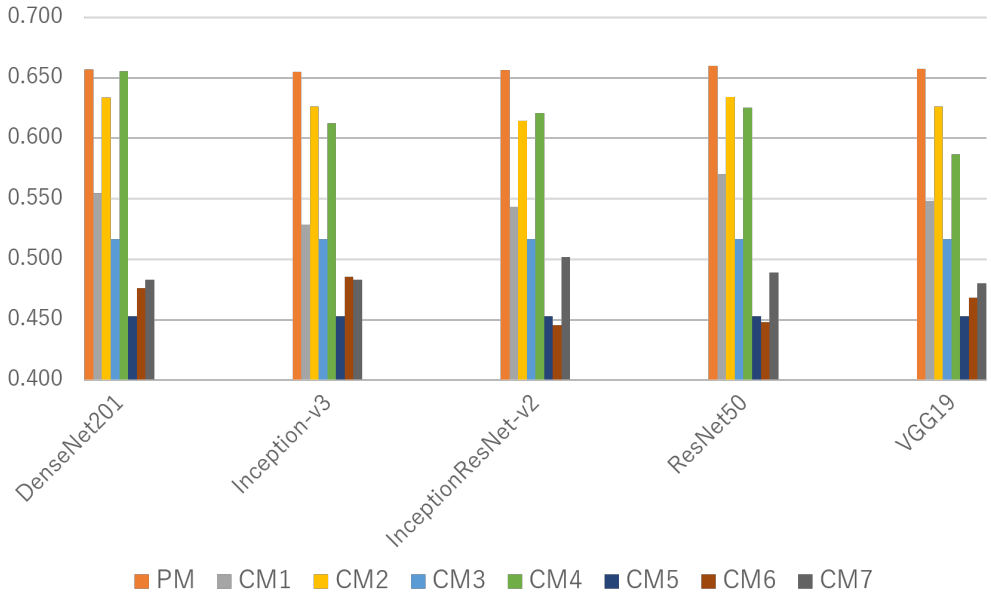

Figure 4.1.3: Average F-measure across all emotion labels and participants for each CNN feature extraction model.

A method [45] employing deep canonical correlation analysis (DeepCCA) [46] for estimating emotion labels based on CNN features and gaze features [97].

- CM7

  This method estimates emotion labels using general feature fusion based on CCA [98] between gaze features [97] and CNN features.

To evaluate the resilience of our approach, we utilized five commonly used CNN models: Inception-v3 [99], DenseNet201 [73], Inception-ResNet-v2 [79], VGG19 [100] and ResNet50 [101] for extracting CNN features.

Table 4.1.1: Average F-measure across all emotion labels and CNN features for each participant.

	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Par10	Par11	Par12	Par13	Average
PM	**0.668**	**0.667**	**0.628**	**0.652**	**0.668**	**0.660**	**0.670**	**0.665**	**0.667**	**0.647**	**0.635**	**0.668**	**0.648**	**0.657**
CM1	0.540	0.533	0.544	0.582	0.565	0.520	0.523	0.571	0.560	0.536	0.529	0.566	0.569	0.549
CM2	0.628	0.623	0.625	0.642	0.628	0.623	0.606	0.625	0.632	0.615	0.626	0.636	0.642	0.627
CM3	0.508	0.564	0.472	0.521	0.555	0.502	0.456	0.534	0.520	0.528	0.502	0.573	0.482	0.517
CM4	0.614	0.641	0.621	0.619	0.619	0.626	0.622	0.640	0.628	0.619	0.601	0.609	0.603	0.620
CM5	0.546	0.371	0.379	0.520	0.488	0.397	0.396	0.473	0.463	0.471	0.468	0.406	0.537	0.453
CM6	0.471	0.337	0.479	0.450	0.486	0.436	0.489	0.462	0.460	0.439	0.480	0.493	0.511	0.476
CM7	0.499	0.475	0.449	0.496	0.492	0.479	0.473	0.449	0.459	0.504	0.525	0.481	0.515	0.485

Table 4.1.2: Average F-measure across all participants and CNN features for each emotion label.

	Amusement	*Anger*	*Awe*	*Content*	*Disgust*	*Excitement*	*Fear*	*Sad*	Average
PM	0.659	0.594	**0.663**	**0.667**	**0.667**	**0.668**	0.672	**0.667**	**0.657**
CM1	0.625	0.626	0.479	0.538	0.525	0.506	0.553	0.541	0.549
CM2	0.671	**0.659**	0.604	0.607	0.631	0.600	0.630	0.614	0.627
CM3	0.448	0.533	0.461	0.595	0.572	0.420	0.622	0.482	0.517
CM4	**0.781**	0.557	0.527	0.627	0.594	0.566	**0.683**	0.625	0.620
CM5	0.418	0.449	0.453	0.498	0.435	0.494	0.441	0.435	0.453
CM6	0.465	0.460	0.463	0.455	0.486	0.446	0.472	0.471	0.476
CM7	0.469	0.455	0.511	0.496	0.485	0.488	0.503	0.493	0.488

The experimental results are presented in Tables 4.1.1 and 4.1.2, along with Fig. 4.1.3. Table 4.1.1 displays the average F-measure across all emotion labels and CNN features for each participant, while Table 4.1.2 displays the results across all participants and CNN features for each emotion label. Furthermore, Fig. 4.1.3 depicts the average F-measure of all participants and emotion labels concerning each CNN feature extraction model. The experimental results consistently show that the PM achieves a higher average F-measure compared to all comparative methods. This validates the effectiveness of our approach. Specifically, when comparing PM with CM1, the effectiveness of our novel image representation, GIT, for emotion label estimation is evident. The comparison with CM2 demonstrates the overall effectiveness of our network architecture for accurate estimation. Comparisons with CMs 3 and 4 highlight the superiority of using both neural networks over individual networks. Additionally, the comparison with CM5 indicates that CNN features outperform hand-crafted features for emotion label estimation. PM also outperforms the CM6 in terms of accuracy. Finally, the effectiveness of our method surpasses that of the simple feature fusion method (CM7) relying on CCA. Furthermore, the robustness of our method is evident, as Fig. 4.1.3 shows consistently higher F-measure values for PM across all CNN features compared to other comparative methods.

4.1.4 Conclusions

This chapter presents the method for estimating emotion labels using tensor-based spatiotemporal visual attention analysis. In order to improve the performance of estimating emotion labels, we consider the temporal dynamics of visual attention in human gaze towards objects within a target image by constructing a fourth-order GIT. Leveraging the generated GIT, two networks are established to independently estimate emotion labels based on temporal changes in visual attention and the presence of objects in the target image. Consequently, our approach achieves emotion label estimations through decision-level fusion of the outputs from these networks.

Chapter 4.2

Emotional Category Classification Using Visual Attention-based Heterogeneous CNN Feature Fusion Based on Tensor Analysis

4.2.1 Introduction

Chapter 4.1 introduces the emotional category classification method based on GIT, which incorporates information on the viewed object and the temporal changes in visual attention. This method collaboratively uses CNN features extracted from GIT and GIT itself for considering both semantic information about objects and the temporal changes in visual attention. While CNN features excel in the ability representing the source domain, they might lack the capability for our target domain. To enhance semantic features and improve representation for emotional category classification, utilizing multiple CNN features from various CNN models is desirable. This requires a heterogeneous feature fusion method that takes into account both temporal changes and interactions among CNN features. Due to the high dimensionality of CNN features, their fusion and analysis present challenges. Therefore, our emphasis is on a tensor-based feature fusion approach, resembling vector concatenation. The dimensions of each mode in the formed tensor are less than that of vector concatenation. Consequently, utilizing tensor-based feature fusion allows for the examination of temporal changes and interactions among CNN features. However, managing high-order information, encompassing specifics about CNN features, their quantity, and temporal changes, is essential. Hence, for emotional category clas-

sification, it is imperative to employ a learning method that incorporates tensor analysis.

In this chapter, we introduce a novel approach for tensor-based emotional category classification, employing visual attention-based heterogeneous CNN feature fusion. Multiple CNN features are extracted from each frame of GIT, where frame in our proposed method refers to the pairing of the image and visual attention at each time unit that divides the total gaze time in GIT. It is worth mentioning that while frame typically refers to a unit in a movie, in our context, it denotes this unique pairing. Additionally, several CNN features are extracted and used to construct a new CNN feature-based tensor (CFT) to consider the interactions between CNN features. Given that each feature of CFT originates from GIT, we anticipate that the proposed method enables visual attention-based heterogeneous CNN feature fusion, ultimately enhancing representation ability. The primary contribution of this chapter lies in CNN feature fusion based on CFT. Finally, we leverage General Tensor Discriminant Analysis (GTDA) [92], which is supervised feature transformation specific to the tensor analysis, for CFT. GTDA transforms input features into more highly discriminative ones, facilitating emotional category classification through Logistic Tensor Regression (LTR) [102]. It is noteworthy that both GTDA and LTR are tensor-based methods and are applied for analyzing CFTs. Consequently, precise classification of emotional categories is attainable through this novel feature fusion approach.

4.2.2 Tensor-based Emotional Category Classification

The proposed method classifies images into emotional categories through tensor-based analysis, which realizes heterogeneous feature fusion based on visual attention, specifically tailored to our target problem. The flow of our approach is presented in Fig. 4.2.1. Details regarding CNN feature extraction and the construction of the CFT can be found in Section 4.2.2.1. Emotional category classification using the transformed CFT via GTDA and LTR are covered in Section 4.2.2.2, respectively. Given the confirmed effectiveness of combining GTDA and LTR in [93], we have incorporated these techniques into our method.

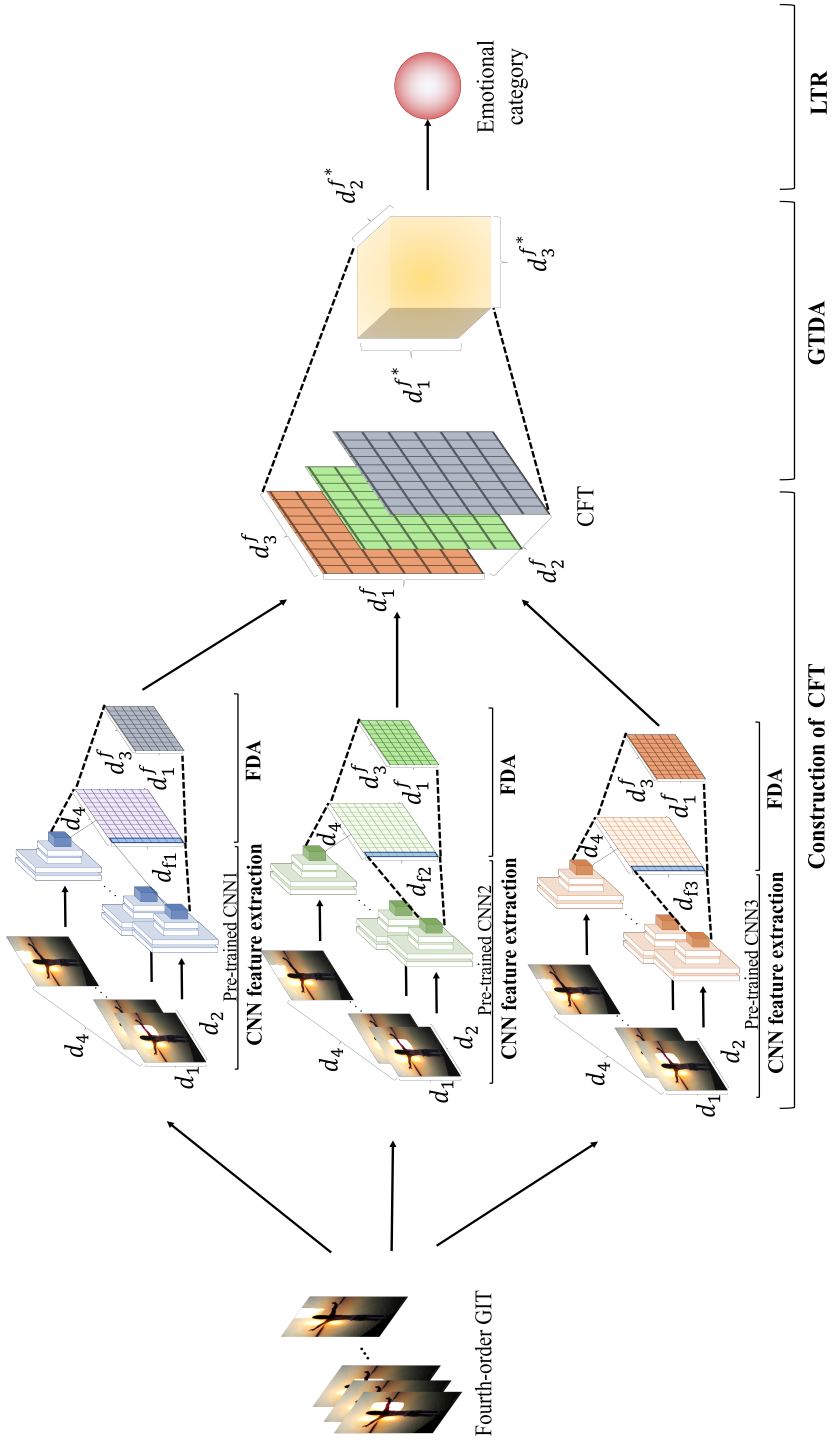

Figure 4.2.1: Flow of our approach. We establish a gaze-based image representation and extract various types of CNN features. Through the alignment of these CNN features, we create a CFT and apply both GTDA and LTR to this tensor. At the end, our method performs image classification into emotional categories utilizing the outputs generated by the proposed network.

4.2.2.1 CNN Feature Extraction and CFT Construction

The proposed method involves the extraction of CNN features by using the outputs of the last pooling layer of pre-trained CNNs. Specifically, three CNN models, namely DenseNet201 [73], Inception-ResNet-v2 [79], and Xception [103], are utilized for feature extraction. The dimensions of these CNN features are 1920, 1536, and 2048, respectively. In our approach, a CFT is constructed by aligning these features extracted from the GIT presented in Section 4.1.2.1. However, due to the disparate dimensions of these CNN features, direct spatial concatenation is challenging. To address this, we employ supervised dimension reduction, specifically Fisher discriminant analysis (FDA) [104], to unify their dimensions to the lowest one, i.e., 1536. Consequently, by aligning these CNN features, our method forms the CFT $\mathcal{V}_n^{\mathrm{3rd}} \in \mathbb{R}^{d_1^f \times d_2^f \times d_3^f}$, where d_1^f represents the minimum CNN feature dimension (1536), d_2^f signifies the number of CNN features (three), and d_3^f denotes the number of frames, equal to d_4.

Incorporating multiple CNN features in our method enhances the representation ability, and the CFT introduces a novel way of considering CNN feature dimensions, temporal changes in visual attention, and the types of CNN features. This enables simultaneous consideration of interactions among multiple types of CNN features, resulting in heterogeneous CNN feature fusion for enhancing the representation ability.

4.2.2.2 LTR-based Emotional Category Classification

As the input of the LTR-based classifier, we prepare the transformed CFT $\hat{\mathcal{V}}_n^{\mathrm{3rd}}$, which is computed by applying GTDA presented in Section 4.1.2.3 to the CFT $\mathcal{V}_n^{\mathrm{3rd}}$. For the CFT $\hat{\mathcal{V}}_{\mathrm{test}}^{\mathrm{3rd}} \in \mathbb{R}^{d_1^{f^*} \times d_2^{f^*} \times d_3^{f^*}}$ calculated from the test image, we aim to predict its class label y_{test}. The formulation of the LTR model employed in our method is as follows:

$$\Pr[y_{\mathrm{test}} \mid \hat{\mathcal{V}}_{\mathrm{test}}^{\mathrm{3rd}}, \mathcal{Z}] = \frac{1}{1 + \exp(-\langle \mathcal{Z}, \hat{\mathcal{V}}_{\mathrm{test}}^{\mathrm{3rd}} \rangle)}, \tag{4.2.1}$$

where $\mathcal{Z}$ represents a parameter tensor containing regression coefficients, and it shares the same size as the transformed CFT $\hat{\mathcal{V}}_n$. To determine the optimal parameter tensor $\hat{\mathcal{Z}}$ for $\mathcal{Z}$, we address

the following maximum log-likelihood problem:

$$\hat{\mathcal{Z}} = \arg\max_{\mathcal{Z}} \mathcal{L}(\mathcal{Z}), \tag{4.2.2}$$

where

$$\mathcal{L}(\mathcal{Z}) = \sum_{n=1}^{N} (y_n \ln\left(\langle \mathcal{Z}, \hat{\mathcal{V}}_n^{3\text{rd}} \rangle\right) + (1 - y_n) \ln\left(1 - \langle \mathcal{Z}, \hat{\mathcal{V}}_n^{3\text{rd}} \rangle\right)). \tag{4.2.3}$$

The maximization problem mentioned above can be addressed by incorporating an L_1-norm regularization term for $\mathcal{Z}$, inspired by the approach in [102].

Finally, the proposed method predicts the class label as follows:

$$y_{\text{test}} = \arg\max_{y \in \{0,1\}} \Pr[y \mid \hat{\mathcal{V}}_{\text{test}}^{3\text{rd}}, \hat{\mathcal{Z}}]. \tag{4.2.4}$$

Therefore, the proposed approach achieves heterogeneous CNN feature fusion and tensor-based analysis while taking into account temporal changes in visual attention.

4.2.3 Experiments

4.2.3.1 Experimental Conditions

This section presents the experimental results validating the effectiveness of our proposed method. The experiments utilized the abstract paintings dataset [39], comprising 280 images, and each image is labeled with at least one emotion from eight categories (*amusement, awe, contentment, excitement, anger, disgust, fear*, and *sad*). The ground truth (GT) for each image was derived from emotion labels assigned by around 14 individuals. Our method was applied to estimate each emotion label, using 224 randomly selected images for training and remaining 56 images for testing. The performance evaluation employed F-measure, the harmonic mean of Recall and Precision. To carefully consider the data imbalance, the number of images for each class was equalized through random selection and the proposed method was trained for each emotion label. Tobii Eye Tracker 4C[1] was used in this experiment with 13 participating

[1] https://tobiigaming.com/eye-tracker-4c/

participants (Pars 1-13). Participants were tasked with viewing each image until recalling some emotions, with one second allocated to adjust their gaze to the center of the monitor before viewing each image. The length of gazing time was normalized as it became d_4.

For comparing the proposed method (PM), we employed eight comparative methods (CMs). CM1 excluded the utilization of temporal changes in visual attention from the PM. Therefore, in CM1, $d_4 = 1$ in the GIT. Additionally, CM1 employed only one CNN feature among the three types presented in Section 4.2.2.1. CM1 was incorporated to assess the innovative approaches introduced in this chapter. CM2 utilized solely gaze features extracted based on [97], classifying emotional categories using an Extreme Learning Machine (ELM) [94]. We adopted CM2 to evaluate the incorporation of both gaze information and image through the comparison of the proposed method. Moreover, we compared with the following three methods. First, we adopted CM3 [47] that employed both hand-crafted visual features and gaze information. Besides, since multi-modal features were used in the experiment, this fusion method was considered suitable for comparison. An emotional category classification method, which fuses multiple types of biological data through Deep Canonical Correlation Analysis (Deep CCA) [46], was proposed [45], and we employed this method as CM4 with gaze features [97] and CNN features. Moreover, we employed CM5 that performs the image classification into emotional categories through CCA [98]-based feature fusion applied to both gaze features [97] and CNN features. CMs 6 and 7 employed CNN feature fusion based on vector concatenation. Specifically, CM6 constructed a two-order CFT with dimensions corresponding to CNN features and their changes over time. CM6 concatenated multiple CNN features at each time, applying GTDA to the formed two-order CFT. On the other hand, CM7 concatenated all CNN features, treating the vector whose dimension is the product of the dimension of CNN features and the number of CNN feature types. To prevent an increase in dimensionality, CM7 averages the temporal changes in CNN features. Handling the resulting vector, CM7 used linear discriminant analysis (LDA) [104] in the place of GTDA. At the end of CMs 6 and 7, Support Vector Machine (SVM) [105] and ELM were applied to obtained features for classifying input images into emotional categories. While, CM8 fused CNN features based on decision-level feature fusion. Concretely, in CM8, we initially constructed a two-order CFT comprising CNN features with considering their temporal changes in each CNN features. Then GTDA were applied for feature transformation of two-order CFT,

and the subsequent ELM or SVM classified transformed two-order CFT into the emotional category. Notably, the decision-level feature fusion technique were employed for treating multiple modalities in CM8 [106, 107].

Table 4.2.1: Mean F1-measure scores across all emotional categories computed for each participant. It is worth noting that •-•-• and •,•,• differ in the consideration of their order.

	PM	PM	PM	CM1	CM1	CM1	CM2 [97]	CM3 [47]	CM4 [45]	CM5 [98]	CM6	CM6	CM7	CM7	CM7	CM8	CM8
CNN Feature	D-I-X	D-X-I	X-D-I	D	I	X	-	-	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X
Classifier	ELM	ELM	ELM	ELM	ELM	ELM	ELM	SVM	SVM	SVM	SVM	ELM	SVM	ELM	SVM	SVM	ELM
Par1	0.616	0.613	**0.631**	0.426	0.468	0.452	0.567	0.546	0.512	0.529	0.533	0.543	0.578	0.544	0.531	0.560	
Par2	0.616	0.591	**0.635**	0.514	0.429	0.486	0.402	0.371	0.311	0.455	0.508	0.519	0.413	0.414	0.506	0.570	
Par3	**0.592**	0.552	0.538	0.469	0.415	0.362	0.490	0.379	0.567	0.538	0.531	0.567	0.442	0.501	0.494	0.498	
Par4	0.567	0.558	**0.583**	0.401	0.417	0.423	0.402	0.520	0.532	0.540	0.502	0.563	0.503	0.571	0.468	0.527	
Par5	**0.606**	0.603	0.551	0.507	0.446	0.441	0.512	0.488	0.564	0.502	0.489	0.560	0.504	0.525	0.473	0.528	
Par6	**0.603**	0.542	0.589	0.486	0.453	0.440	0.505	0.397	0.478	0.512	0.495	0.526	0.504	0.498	0.446	0.528	
Par7	0.565	**0.643**	0.593	0.393	0.498	0.341	0.521	0.396	0.440	0.489	0.487	0.592	0.513	0.507	0.469	0.546	
Par8	0.567	0.597	**0.598**	0.378	0.424	0.533	0.468	0.473	0.567	0.537	0.510	0.511	0.565	0.588	0.400	0.515	
Par9	**0.598**	0.558	0.597	0.471	0.436	0.414	0.505	0.463	0.565	0.479	0.503	0.475	0.498	0.433	0.484	0.539	
Par10	0.560	0.526	**0.603**	0.463	0.395	0.400	0.425	0.471	0.522	0.563	0.487	0.483	0.447	0.474	0.451	0.555	
Par11	**0.628**	0.564	0.568	0.465	0.431	0.433	0.423	0.468	0.468	0.494	0.548	0.521	0.497	0.492	0.408	0.599	
Par12	**0.597**	0.597	0.588	0.333	0.442	0.408	0.414	0.406	0.497	0.466	0.509	0.515	0.562	0.529	0.489	**0.625**	
Par13	0.570	0.579	**0.598**	0.503	0.487	0.381	0.667	0.537	0.409	0.495	0.550	0.537	0.436	0.440	0.506	0.584	
Average	**0.591**	0.579	0.590	0.447	0.442	0.424	0.485	0.453	0.487	0.502	0.512	0.532	0.497	0.503	0.471	0.552	

Table 4.2.2: Mean F1-measure scores across all participants computed for each emotional category. It is worth noting that •-•-• and •,•,• differ in the consideration of their order.

	PM	PM	PM	CM1	CM1	CM1	CM2 [97]	CM3 [47]	CM4 [45]
CNN Feature	D-I-X	D-X-I	X-D-I	D	I	X	-	-	D,I,X
Classifier	ELM	ELM	ELM	ELM	ELM	ELM	ELM	SVM	SVM
Amusement	**0.667**	**0.667**	**0.667**	0.409	0.473	0.564	0.527	0.418	0.531
Anger	**0.667**	**0.667**	**0.667**	0.605	0.506	0.446	0.527	0.449	0.493
Awe	**0.667**	**0.667**	**0.667**	0.360	0.354	0.410	0.529	0.453	0.469
Content	0.424	0.414	0.394	0.426	0.443	0.355	0.411	0.498	0.497
Disgust	**0.667**	**0.667**	**0.667**	0.452	0.481	0.475	0.521	0.435	0.505
Excitement	0.550	0.599	**0.630**	0.431	0.450	0.419	0.434	0.494	0.475
Fear	0.612	0.523	0.563	0.452	0.388	0.366	0.531	0.441	0.456
Sad	0.474	0.426	0.468	0.438	0.439	0.358	0.402	0.435	0.473
Average	**0.591**	0.579	0.590	0.447	0.442	0.424	0.485	0.453	0.487

	CM5 [98]	CM6	CM6	CM7	CM7	CM8	CM8
CNN Feature	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X	D,I,X
Classifier	SVM	SVM	ELM	SVM	ELM	SVM	ELM
Amusement	0.501	0.486	0.648	0.488	0.489	0.462	0.622
Anger	0.517	0.502	0.493	0.505	0.545	0.483	0.490
Awe	0.536	0.387	0.500	0.648	0.457	0.540	0.526
Content	0.503	**0.553**	0.501	0.502	0.514	0.389	0.519
Disgust	0.476	0.420	0.488	0.460	0.450	0.524	0.481
Excitement	0.500	0.527	0.533	0.480	0.482	0.544	0.622
Fear	0.495	**0.660**	0.547	0.530	0.551	0.406	0.551
Sad	0.487	0.558	0.544	0.543	0.533	0.421	**0.602**
Average	0.502	0.512	0.532	0.497	0.503	0.471	0.552

4.2.3.2 Performance Evaluation

Tables 4.2.1 and 4.2.2 present the outcomes of the experiment. Table 4.2.1 displays the average F1-measures for all emotional categories, computed for each participant. Table 4.2.2 displays the average F1-measure computed for each emotional category across all participants. ”D,” ”I,” and ”X” denote DenseNet201, Inception-ResNet-v2, and Xception, respectively. The order in which CNN features are combined influences emotional category estimation performance, and comparison of PM (D-I-X), PM (D-X-I), and PM (X-D-I) reveals that PM (D-I-X) yields the best results on average, influenced by the mode expansion in the second mode of GTDA within our method. Despite PM (D-X-I) exhibiting the least favorable results among PMs, it outperforms all comparative methods, affirming the effectiveness of PM without considering the combination order of CNN features. The decision method for this order possesses intriguing characteristics that warrant future consideration, although our current focus is on heterogeneous CNN feature fusion and analysis.

The proposed method outperforms comparative methods based on the obtained results. Comparison of PM with CM1 validates the effectiveness of novel approaches adopted in our method. A comparison of PM with CMs 1 and 2 affirms the efficacy of the new gaze-based image representation and CFT, demonstrating the benefits of collaboratively using image and gaze information. PM surpasses CM3 and CM4 in F1-measure, which shows the superior performance of PM in classifying images into emotional categories. Comparison with CM5 highlights the effectiveness of combining gaze information and images using both the new gaze-based image representation and CFT, outperforming baseline fusion methods. Additionally, PM excels over CMs 6, 7, and 8, emphasizing the superiority of our proposed heterogeneous CNN feature fusion and its analysis over vector-based concatenation methods for emotional category classification.

In addition to quantitative evaluations, a representative experimental results are depicted in Fig. 4.2.2. The gaze-based image representations of Par2 and Par7 are classified into four categories, encompassing all ground truths, whereas that of Par8 is classified into three categories, covering one ground truth. Par2 and Par7 viewed nearly identical areas in each frame of the shown image, while Par8 viewed a different area, resulting in varying classified emotional categories. This observation confirms the relationship between temporal changes in visual attention

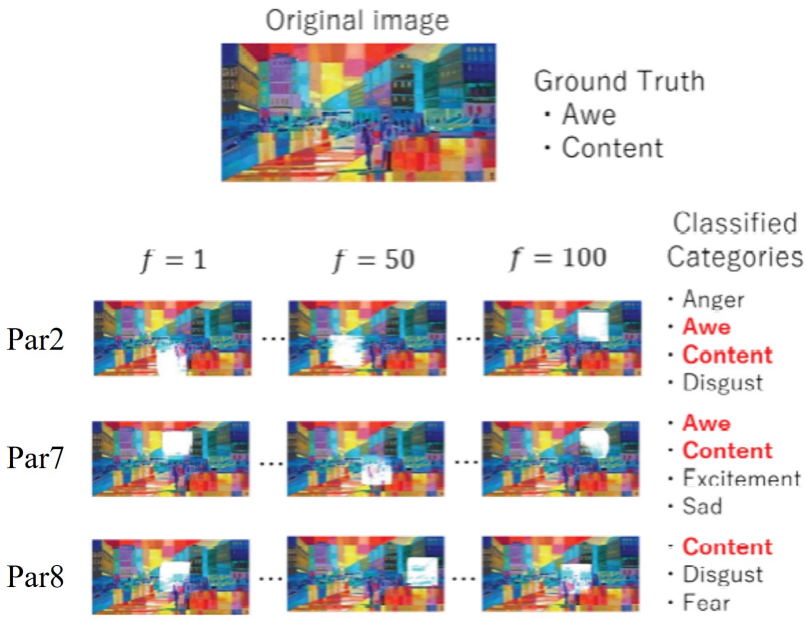

Figure 4.2.2: Selected experimental outcomes. This figure depicts a set of test images alongside their corresponding ground truths. The regions where participants viewed are highlighted in white at frames 1, 50, and 100. Utilizing gaze data, PM (D-I-X) assigns categories to the image. If the assigned category matches the ground truth, the corresponding category is denoted in red.

and human emotions.

4.2.4 Conclusions and Discussions

In this chapter, we have introduced a tensor analysis-based method for emotional category classification, achieving visual attention-based heterogeneous CNN feature fusion. To enhance classification performance, our method introduces a new tensor, CFT, which consolidates outputs from multiple CNN models while considering temporal changes in visual attention. Subsequently, emotional category classification is made possible through the application of GTDA and LTR. The effectiveness of our proposed method was confirmed through experimental results.

Chapter 5

Multi-modal Human Emotion Recognition

Human emotions, an essential yet enigmatic aspect of human nature, hold the potential to contribute to various fields. One such area is multimedia content recommendation, where understanding the mechanism underlying the occurrence of human emotions facilitates personalized preferences in recommendations [108, 109]. Additionally, in the field of human-computer interaction, implementing this mechanism allows agents in robots or computers to engage affectively with humans [110, 111]. The pursuit of integrating this mechanism into computers is referred to as affective computing [112]. Recognizing human emotions is a crucial initial step in affective computing, but it remains challenging due to the subjective nature of human emotions.

In the domain of signal processing, there have been investigations into examining brain activities recorded during humans viewing images/videos or listening to music as a means of recognizing human emotions [93, 113]. However, emotions derived from such brain activity analysis may not necessarily correlate with the target stimuli, given the vast amount of information processed by the human brain from various sources. Conversely, non-verbal cues such as facial expressions and eye gaze have been explored as indices for the recognition of human emotions [13]. These non-verbal cues have the potential to encapsulate subconscious reactions, governed by the sympathetic nervous system irrespective of human intention. In this way, incorporating these cues alongside brain activity enhances the capture of more reliable emotion-related information. Indeed, multi-modal human emotion recognition methods, utilizing multiple types of biological

data, have demonstrated superior accuracy compared to uni-modal methods [45, 48, 50, 114]. The majority of these multi-modal approaches leverage eye gaze and brain activity as modalities to capture explicit and implicit information, respectively.

Previous approaches have involved the cooperative utilization of multiple types of biological data by integrating features computed from such data, prompting researchers to explore more effective methods for feature integration conducive to emotion analysis. Notably, in [48] and [45], Bi-modal Deep Autoencoder (BDAE) [49] and Deep Canonical Correlation Analysis (DeepCCA) [46] are employed to extract common factors across all features. Despite the sequential nature of biological data, these methods overlook the temporal changes in biological data which are crucial aspects of human emotion recognition [77, 115]. Subsequently, the study [50] introduces Bi-modal Long-Short Term Memory (BLSTM) [51] to consider temporal changes, successfully capturing temporal dynamics by aligning each modality at the same timestep. However, these methods adopt general machine learning frameworks, neglecting the intrinsic properties of biological data. In this chapter, we strive to develop several machine learning models specific to the characteristics of biological data.

Chapter 5.1

Human-centric Emotion Estimation Based on Correlation Maximization Considering Changes with Time in Visual Attention and Brain Activity

5.1.1 Introduction

When utilizing biological data, the consideration of user burden in data acquisition is paramount. While gaze data can be obtained using small sensors like those in glasses, acquiring brain activity data still poses a significant user burden. Moreover, studies have indicated the relevance of temporal changes in visual attention and gazed objects to human emotion [77] when using gaze data. Although some studies have attempted to estimate human emotions based on gaze and brain activity data [45,48,50,116], these efforts did not specifically address the temporal changes in biological data. Therefore, achieving higher-performance emotion estimation necessitates the collaborative use of temporal changes in both gaze data and brain activity data.

Based on the above considerations, this chapter addresses the following two problems:

1. To alleviate user burden, brain activity data is obtained only during the training phase. In other words, a method that collaboratively utilizes gaze and brain activity data without requiring brain activity data acquisition during the test phase is sought.

2. The temporal changes in both gaze data and brain activity data need to be considered to enhance emotion estimation accuracy. Brain activity data with higher temporal resolution

is preferable, and analyzing the relationship between gaze and brain activity data at each timestep is expected to improve accuracy.

This chapter introduces human-centric emotion estimation that maximizes correlation between visual attention and brain activity, considering temporal changes. The term "human-centric" is used since the proposed method is trained individually for each user, focusing on extracting their implicit states. Canonical Correlation Analysis (CCA) [98] is employed to address the above problems, offering the following solutions:

1. Transformation Matrix Calculation

   CCA computes a transformation matrix from two types of features. Once calculated, this matrix remains constant, eliminating the need for recalculation. As brain activity data is only required during the training phase, this reduces the user burden.

2. Use of Gaze and Image Tensor (GIT) for Multi-modal Emotion Recognition

   A GIT [117] is constructed to represent temporal changes in visual attention as presented in Section 4.1.2.1. This innovative approach incorporates time as an axis corresponding to the frame in addition to the axis of images, providing a novel image representation considering temporal changes in visual attention.

The efficacy of CCA has been extensively reported across various domains, including computer vision and human-computer interaction [118–121]. In this way, we employ the CCA-based approach to integrate gaze-based visual features and brain activity-based features. Initially, for concurrent analysis of images and temporal changes in visual attention, we utilize the fourth-order GIT [117]. The first and second modes of this tensor represent pixel locations, and the third mode corresponds to color channels, encapsulating image information. Additionally, the fourth mode of the tensor considers temporal changes, corresponding to frames. Moreover, by feeding the acquired GIT into Convolutional Neural Network (CNN) [4] models, our method facilitates the derivation of novel gaze-based visual features. Subsequently, our method transforms these gaze-based visual features to acquire emotion-correlated features by maximizing canonical correlation through CCA, utilizing brain activity-based features obtained from users viewing images. Through these feature extraction and transformation methods, our approach derives human-centric features tailored for emotion estimation. Another advantage lies in the fact

that brain activity-based features are only used for obtaining feature transformation, and making their acquisition unnecessary for estimating emotion from newly obtained images indicates the broad applicability. Finally, in the classification step, our method derives human-centric visual features from multiple CNN models, yielding a third-order tensor with modes corresponding to "dimensions of the transformed features," "types of adopted CNN models," and "the time axis." By applying generalized tensor discriminant analysis (GTDA) [92] to this third-order tensor and conducting classification using an extreme learning machine (ELM) [94] capable of training with a limited number of samples, our proposed method achieves human-centric emotion estimation.

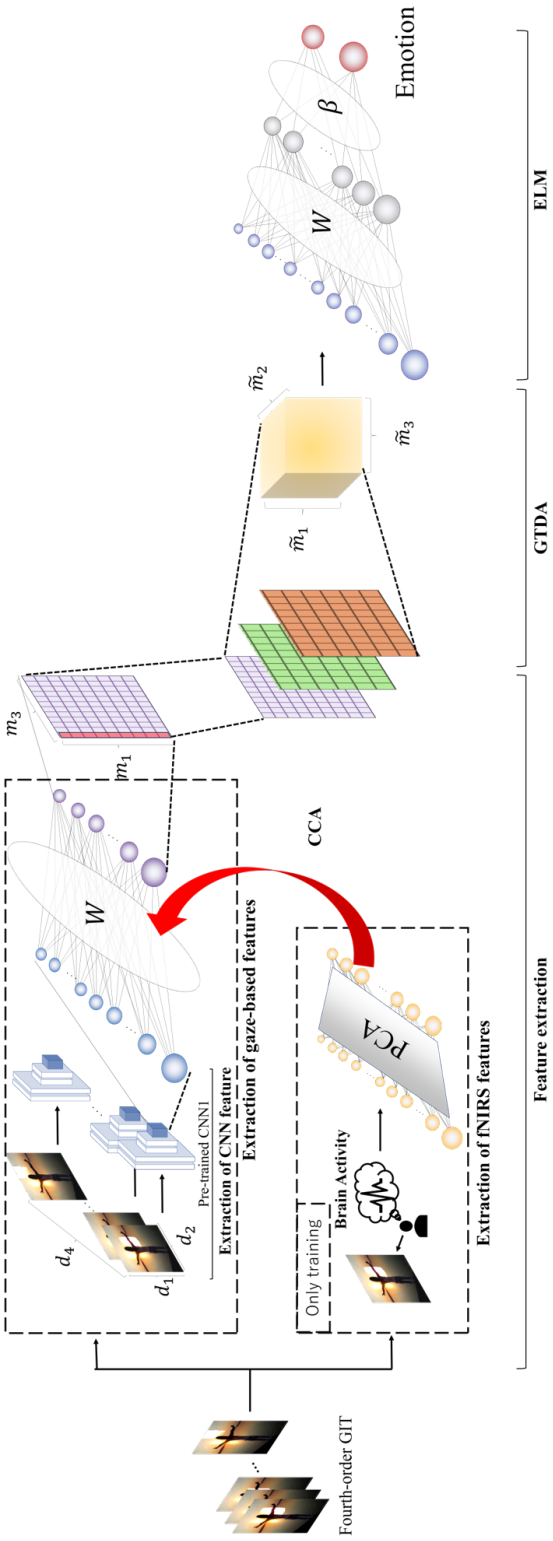

Figure 5.1.1: Flow of the entire model presented in this chapter.

5.1.2 Our Estimation Method

The proposed method comprises three steps, outlined in Fig. 5.1.1. In the initial step, we compute gaze-based visual features. Our method utilizes the fourth-order GIT as an image representation, considering both objects in images and temporal changes in visual attention. Subsequently, we derive pre-trained CNN-based visual features from the images corresponding to each frame of the fourth-order GIT, serving as gaze-based visual features. In the second step, we compute brain activity-based features [18] from each user. Additionally, CCA is applied between gaze-based visual features and brain activity-based features at each frame for transforming gaze-based visual features into novel features that account for temporal variations. In the final step, we align all transformed features and construct a new third-order tensor. Emotion estimation is then performed using tensor-based machine learning.

5.1.2.1 Gaze-based Visual Feature Extraction via GIT

Initially, to analyze images and temporal changes in visual attention simultaneously, we construct a fourth-order GIT $\mathcal{X}_{n,f}^{4\text{th}}$ presented in 4.1.2.1. Subsequently, to extract more semantically meaningful gaze-based visual features, we derive three types of visual features from the image corresponding to each frame of the fourth-order GIT. To enhance the representation of human emotions, we incorporate various types of visual features. As the visual features calculated from the GIT capture objects in images and the temporal changes in visual attention, we consider these visual features as gaze-based visual features. Specifically, we utilize the outputs of an intermediate layer included in several CNN models, given the well-established efficacy of CNNs in object recognition [122]. By constructing the GIT based on CNN features, we obtain gaze-based visual features that characterize objects viewed by humans.

Training CNNs requires a substantial amount of data. However, preparing a such amounts of GITs is challenging due to the limited gaze data from each user. To address this, we employ transfer learning, which is a proven effective technique [123]. Generally, the CNNs are pre-trained using the ImageNet dataset [4]. Our method incorporates three CNN models, namely, Xception (X) [103], InceptionResnet-v2 (I) [79], and Densenet201 (D) [73]. We extract visual

features $\boldsymbol{v}_{n,f}^{p} \in \mathbb{R}^{d_p}$ ($p \in$ {X, I, D}, with d_p denoting the dimension of the outputs obtained from the last pooling layer of the CNN model p) based on the pre-trained CNN from the image corresponding to each frame of the fourth-order GIT $\mathcal{X}_{n,f}^{4\text{th}}$. Therefore, our approach extracts gaze-based visual features with considering objects in images and the temporal changes in visual attention from the new image representation, the fourth-order GIT.

5.1.2.2 Extraction of Brain Activity-based Features and CCA-based Transformation

This section describes on the extraction of brain activity-based features and the CCA-based transformation, taking into account the temporal changes. As the way to obtain the brain activity data, various types of measurements, such as EEG, fMRI, and fNIRS, are available. Then, EEG and fNIRS data are particularly noteworthy for their high temporal resolution. Notably, fNIRS, which measures blood oxygenation changes, is robust against external activities such as eye blinks that may occur during image viewing [124]. Additionally, fNIRS equipment imposes minimal behavioral or physical restrictions on users [125]. Therefore, several studies have explored the relationship between human emotions and fNIRS signals [126–128], and in our study, we incorporate fNIRS signals alongside gaze data. Previous research has also utilized both fNIRS and gaze data [129–131]. To capture fNIRS signals, we measure changes in deoxygenated and oxygenated hemoglobin levels in the head cortex using near-infrared light. Subsequently, we compute fNIRS features from fNIRS signals while users view images, based on the approach outlined in [18]. Specifically, we derive the following 11-dimensional features from each channel in each frame, as illustrated in Fig. 5.1.2.

- Statistical features (six dimensions)

  General statistics, encompassing average, variance, skewness, kurtosis, zero-crossing rate, and root-mean-square, are computed for fNIRS signals in each time domain.

- Wavelet transform [132]-based features (five dimensions)

  Applying discrete wavelet transform to fNIRS signals enables their conversion into a frequency domain, comprising both high-frequency and low-frequency components. Subse-

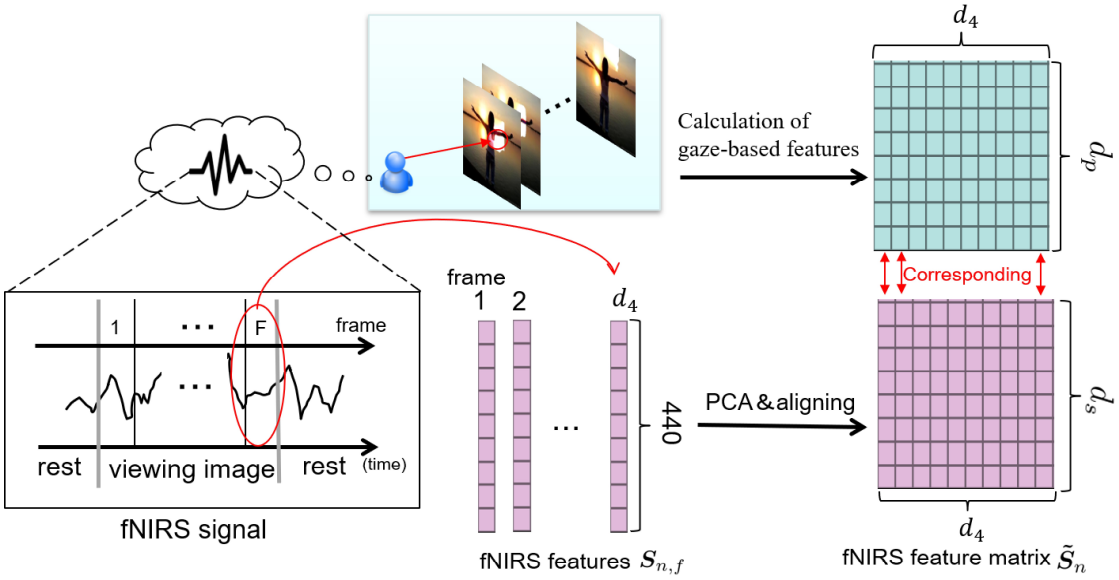

Figure 5.1.2: In our proposed approach, the calculation of brain activity-based features is aligned with gaze-based visual features in each frame.

quently, we determine the energy ratio for each frequency component concerning the total energy.

In the proposed approach, fNIRS signals are acquired from ten channels located on the front and back of the head, respectively. It is important to note that we measure changes in both oxygenated and deoxygenated hemoglobin levels, which provide biologically relevant information for brain function. Consequently, the dimension of the fNIRS features is 440, calculated as 11-dimensional features × 20 channels × 2 (oxygenated/deoxygenated). Consequently, we compute fNIRS features $\boldsymbol{s}_{n,f} \in \mathbb{R}^{440}$ corresponding to a frame f of the nth image. It is essential to apply dimension reduction to fNIRS features $\boldsymbol{s}_{n,f}$ since CCA tends to overfit the training data when the dimension of fNIRS features $\boldsymbol{s}_{n,f}$ exceeds the number of training images. Principal Component Analysis (PCA) [133] is employed as the dimension reduction method, resulting in newly obtained fNIRS features $\tilde{\boldsymbol{s}}_{n,f} \in \mathbb{R}^{d_s}$ (d_s represents the dimension of fNIRS features after applying dimension reduction).

We conduct CCA between the aforementioned fNIRS features $\tilde{\boldsymbol{s}}_{n,f}$ and the gaze-based fea-

tures $\boldsymbol{v}^p_{n,f}$ at each frame f, as illustrated in Fig. 5.1.3. Specifically, we determine the optimal transformation pair $(\hat{\boldsymbol{w}}^p_{s,f}, \hat{\boldsymbol{w}}^p_{v,f}) \in \mathbb{R}^{d_s} \times \mathbb{R}^{d_p}$ by solving the following maximization problem:

$$\max_{(\boldsymbol{w}^p_{s,f}, \boldsymbol{w}^p_{v,f})} \frac{(\boldsymbol{w}^{p\ \top}_{s,f} \boldsymbol{C}^p_{sv,f} \boldsymbol{w}^p_{v,f})}{\sqrt{\boldsymbol{w}^{p\ \top}_{s,f} \boldsymbol{C}_{ss,f} \boldsymbol{w}^p_{s,f}} \sqrt{\boldsymbol{w}^{p\ \top}_{v,f} \boldsymbol{C}^p_{vv,f} \boldsymbol{w}^p_{v,f}}}, \tag{5.1.1}$$

where $^\top$ denotes the transposition operator. Specifically, the variances $\boldsymbol{C}_{ss,f}$, $\boldsymbol{C}^p_{vv,f}$, and the covariance $\boldsymbol{C}^p_{sv,f}$ at frame f are computed as follows:

$$\begin{aligned} \boldsymbol{C}_{ss,f} &= \frac{1}{N} \tilde{\boldsymbol{S}}_f \tilde{\boldsymbol{S}}_f^\top, \\ \boldsymbol{C}^p_{vv,f} &= \frac{1}{N} \boldsymbol{V}^p_f \boldsymbol{V}^{p\top}_f, \\ \boldsymbol{C}^p_{sv,f} &= \frac{1}{N} \tilde{\boldsymbol{S}}_f \boldsymbol{V}^{p\top}_f, \end{aligned} \tag{5.1.2}$$

where

$$\tilde{\boldsymbol{S}}_f = [\tilde{\boldsymbol{s}}_{1,f}, \tilde{\boldsymbol{s}}_{2,f}, \ldots, \tilde{\boldsymbol{s}}_{N,f}], \tag{5.1.3}$$

$$\boldsymbol{V}^p_f = [\boldsymbol{v}^p_{1,f}, \boldsymbol{v}^p_{2,f}, \ldots, \boldsymbol{v}^p_{N,f}]. \tag{5.1.4}$$

Note that $\tilde{\boldsymbol{S}}_f$ and $\boldsymbol{V}^p_f$ are centered in each frame. Furthermore, we can express this maximization problem as follows:

$$\begin{aligned} &(\hat{\boldsymbol{w}}^p_{s,f}, \hat{\boldsymbol{w}}^p_{v,f}) = \arg\max_{(\boldsymbol{w}^p_{s,f}, \boldsymbol{w}^p_{v,f})} \boldsymbol{w}^{p\top}_{s,f} \boldsymbol{C}_{sv,f} \boldsymbol{w}^p_{v,f} \\ &\text{s.t. } \boldsymbol{w}^{p\top}_{s,f} \boldsymbol{C}_{ss,f} \boldsymbol{w}^p_{s,f} = \boldsymbol{w}^{p\top}_{v,f} \boldsymbol{C}_{vv,f} \boldsymbol{w}^p_{v,f} = 1. \end{aligned} \tag{5.1.5}$$

Additionally, we derive the eigenvalue problem utilizing the method of Lagrange multipliers and L1-regularization [134] as follows:

$$\begin{bmatrix} \boldsymbol{O} & \boldsymbol{C}^p_{sv,f} \\ \boldsymbol{C}^{p\ \top}_{sv,f} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^p_{s,f} \\ \boldsymbol{w}^p_{v,f} \end{bmatrix} = \lambda^p_f \begin{bmatrix} \boldsymbol{C}_{ss,f} + \zeta_s \boldsymbol{I}_s & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{C}^p_{vv,f} + \zeta_v \boldsymbol{I}^p_v \end{bmatrix} \begin{bmatrix} \boldsymbol{w}^p_{s,f} \\ \boldsymbol{w}^p_{v,f} \end{bmatrix}, \tag{5.1.6}$$

where λ^p_f represents the Lagrange multiplier, ζ_s and ζ_v are regularization parameters, and $\boldsymbol{I}^p_v$ and $\boldsymbol{I}_s$ denote the identity matrices. The optimal transformation pair $(\hat{\boldsymbol{w}}^p_{s,f}, \hat{\boldsymbol{w}}^p_{v,f})$ is then derived

by solving the eigenvalue problem. The utilization of the valid top d_z ($\leq \min(d_s, d_X, d_D, d_I)$) transformation pairs in each frame results in the acquisition of transformation matrices $\hat{\boldsymbol{W}}^p_{s,f} \in \mathbb{R}^{d_z \times d_s}$ and $\hat{\boldsymbol{W}}^p_{v,f} \in \mathbb{R}^{d_z \times d_p}$.

The gaze-based visual features $\boldsymbol{v}^p_{n,f}$ are transformed using the obtained transformation matrices $\hat{\boldsymbol{W}}^p_{v,f}$ to compute the transformed features at each frame as the following equation:

$$\hat{\boldsymbol{v}}^p_{n,f} = \hat{\boldsymbol{W}}^p_{v,f} \boldsymbol{v}^p_{n,f}. \tag{5.1.7}$$

Through this process, we acquire the transformed gaze-based visual features that take into account the characteristics of fNIRS features. It is important to note that, once the transformation matrices are obtained, we can transform new gaze-based visual features without the need for fNIRS features in the inference phase. This approach offers two significant contributions. Firstly, brain activity data is only required during the training phase, which reduces the burden on users. Secondly, the transformation pair is computed for each frame, which allows us to consider temporal changes in visual attention and brain activity.

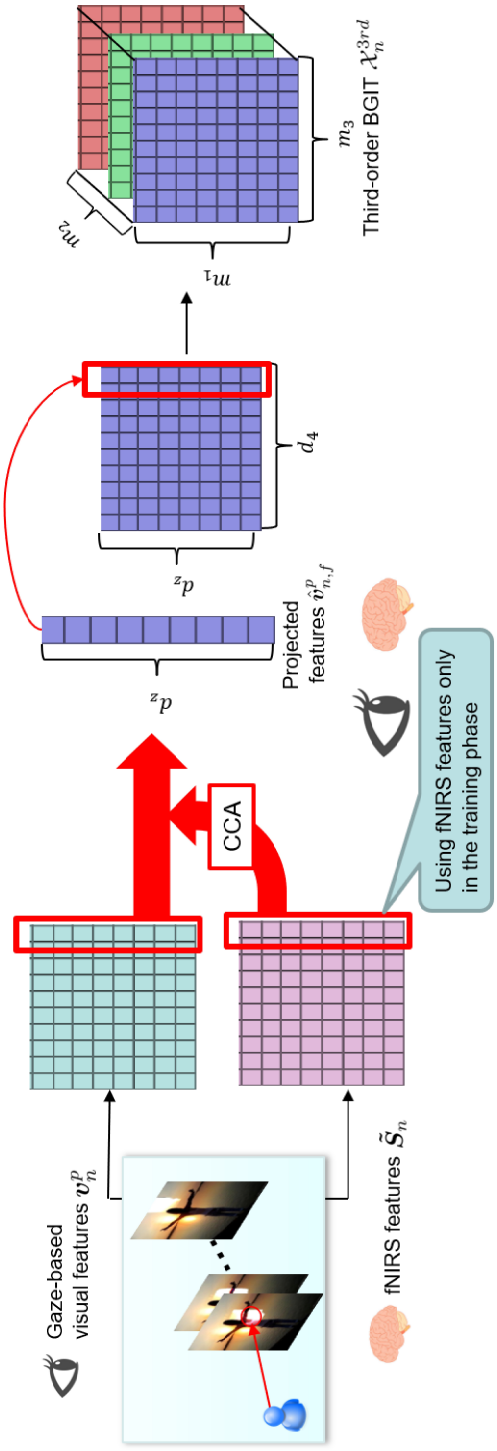

Figure 5.1.3: Feature transformation of gaze-based visual features and fNIRS features based on CCA. fNIRS features are utilized only during the training phase. The per-frame calculation allows for the consideration of temporal changes in visual attention and brain activity.

5.1.2.3 Emotion Estimation Based on Tensor-based Analysis

This section describes emotion estimation based on tensor-based analysis and lightweight machine learning. By uilizing the transformed features $\hat{\mathbf{v}}_{n,f}^{p}$, we construct a new third-order GIT with considering fNIRS (Brain activity-based Gaze and Image Tensor; BGIT) $\mathcal{X}_{n}^{3\mathrm{rd}} \in \mathbb{R}^{m_1 \times m_2 \times m_3}$, to analyze features while taking into account temporal changes. Here, m_1 (= d_z) represents the dimension of transformed features, m_2 (= 3) signifies the types of gaze-based visual features, and m_3 (= d_4) is the number of frames of the fourth-order GIT. Specifically, we construct the third-order BGIT $\mathcal{X}_{n}^{3\mathrm{rd}}$ as follows:

$$\mathcal{X}_{n,i,p,f}^{3\mathrm{rd}} = \hat{\mathbf{v}}_{n,i,f}^{p}, \tag{5.1.8}$$

where $\mathcal{X}_{n,i,p,f}^{3\mathrm{rd}}$ and $\hat{\mathbf{v}}_{n,i,f}^{p}$ represent the (i, p, f) element of the third-order BGIT $\mathcal{X}_{n}^{3\mathrm{rd}}$ and the ith element of the transformed gaze-based visual features $\hat{\mathbf{v}}_{n,f}^{p}$, respectively.

In the proposed approach, we incorporate GTDA and ELM, which are presented respectively in Sections 4.1.2.3 and 4.1.2.4, into the third-order BGIT $\mathcal{X}_{n}^{3\mathrm{rd}}$ to estimate human emotions. Employing GTDA is motivated by its applicability to tensors, and we employ ELM for treating the limited number of training samples. In this way, the proposed method is capable of estimating human emotions by applying tensor analysis and lightweight machine learning to the third-order BGIT $\tilde{\mathcal{X}}_{n}^{3\mathrm{rd}}$.

5.1.3 Experiments

5.1.3.1 Settings

This experiment adopted the Tobii Eye tracker 4C[1] to measure gaze data and the LIGHT-NIRS[2] to measure fNIRS signals. We employed 20 channels, with 10 channels positioned at the front of the head and 10 at the back, as depicted in Fig. 5.1.4. Participants viewed images on a 15-inch display from a distance of 70 cm. Additionally, participants wore a head cap to facilitate the measurement of fNIRS signals, and the gaze sensor was positioned on the display.

[1] https://tobiigaming.com/eye-tracker-4c/
[2] http://www.shimadzu.com/

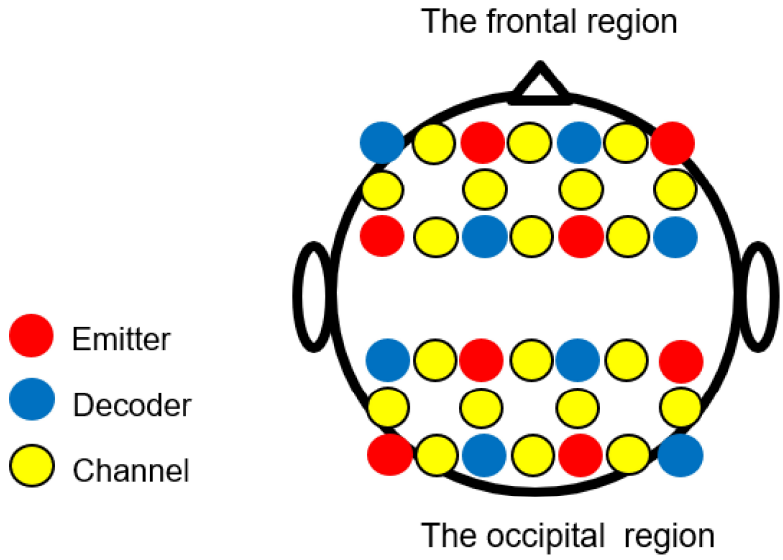

Figure 5.1.4: Channel positions of the measurement instrument for fNIRS signals. We collected fNIRS signals from 20 channels utilizing emitters and decoders.

As the dataset, we utilized the art photo dataset [39]. This dataset comprises images assigned a single label from eight emotional labels (*Amusement, Awe, Contentment, Excitement, Sad, Fear, Anger,* and *Disgust*), and we selected 10 images from images with each emotional label, totaling 80 images. Furthermore, we randomly chose 64 images as training images and used the remaining images as test images.

This experiment involved 10 participants (Pars 1-10) [3]. Participants were instructed to view each image for ten seconds with a ten-second inter-stimulus interval as depicted in Fig. 5.1.5. During the interval, an image with a cross mark in the center was displayed to mitigate the influence of the previous image and guide the gaze to the center of the monitor. Following the task, participants provided feedback (positive/negative) as ground truths regarding the emotions induced by viewing the images. Two emotional states were adopted to increase the number of images in each category. Table 5.1.1 presents the number of samples for each emotion recalled by each participant. It was confirmed that there was no significant difference in the numbers of emotions between participants. To assess the effectiveness of the proposed method, we employed eight comparative methods (CMs 1-8) as follows:

[3]This human research was conducted with the approval of the ethical committee at Hokkaido University.

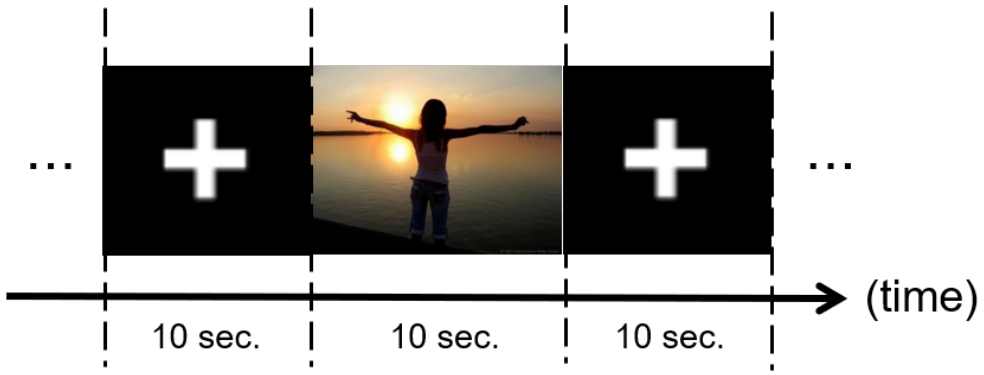

Figure 5.1.5: Experimental design in our experiment. Participants were instructed to view each image for a duration of ten seconds with an inter-stimulus interval of ten seconds. Subsequently, we collected gaze data and fNIRS data from each participant.

Table 5.1.1: Numbers of emotions for participants.

	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Par10
Training Image (Positive)	29	28	30	28	36	23	31	35	28	39
Training Image (Negative)	35	37	34	36	28	41	33	29	36	25
Test Image (Positive)	8	7	8	8	9	7	8	7	7	7
Test Image (Negative)	8	9	8	8	7	9	8	9	9	9

CM1. This approach is akin to the proposed method but employs only a single CNN feature as opposed to multiple types of CNN features.

CM2. This approach estimates emotions solely based on the gaze information used in our method.

CM3. This approach utilizes only fNIRS features in our method, and the ELM-based classifier is used for emotion estimation with fNIRS features.

CM4. This approach handles two modalities, gaze and fNIRS features. It adopts gaze features [97] different from the proposed method. Additionally, the fNIRS features [18] are extracted in the same way as in our method. Two ELM-based classifiers corresponding to each modality are built, and emotions are estimated based on late fusion. It is worth noting that late fusion is a prevalent fusion method, and numerous researchers have employed this method in multimodal analysis [135–137].

In addition to the aforementioned comparative methods, which incorporate elements of our

method, we adopted the following methods for comparing our method.

CM5. This method for emotion estimation [45] utilizes Deep CCA [46] to analyze the relationship between gaze and brain activity information. Three types of CNN features obtained from the GIT are used as gaze features, and Deep CCA is applied to each CNN feature and fNIRS features.

CM6. This emotion estimation method [48] employs a Bi-modal Deep Autoencoder (BDAE) to reconstruct inputs comprising fNIRS and gaze features, enabling the extraction of combined high-level features. It should be emphasized that the calculation of fNIRS and gaze features aligns with our method.

CM7. An extension of CM6, this approach for estimating emotions [116] conducts multi-layer perceptron-based regression from gaze features to the combined high-level features derived from CM6.

CM8. This method for estimating emotions [50] utilizes Long Short-Term Memory (LSTM) [51] for processing gaze and fNIRS features, extracting combined high-level features while accounting for temporal changes. It should be noted that the calculation of gaze and fNIRS features aligns with our method.

As the final classifier, we opted for Support Vector Machine (SVM) [105] over ELM in CMs 5-8, influenced by prior studies [45, 48, 50, 116]. For CMs 5-7, we computed features without considering temporal changes ($d_4 = 1$), as these methods lack the capability to handle such changes. In this experiment, we utilized the F1-measure as an evaluation metric:

Table 5.1.2: The average values are calculated for each participant. The overall average and its standard deviation across all participants are also presented.

	Par1	Par2	Par3	Par4	Par5	Par6	Par7	Par8	Par9	Par10	Average ± std
Ours (I-D-X)	0.80	0.71	0.82	0.71	0.77	0.63	**0.80**	0.84	0.75	0.71	0.76 ± 0.06
Ours (I-X-D)	**0.88**	**0.80**	0.71	**0.80**	0.77	0.63	0.75	0.86	**0.88**	0.67	**0.77 ± 0.08**
Ours (X-I-D)	0.82	0.62	**0.88**	0.62	**0.88**	0.56	0.77	**0.89**	**0.88**	0.71	0.76 ± 0.12
CM1 (X)	0.78	0.62	0.74	0.50	0.63	0.59	0.50	0.74	0.59	0.35	0.60 ± 0.12
CM1 (I)	0.67	0.59	0.70	0.46	0.67	0.53	0.57	0.74	0.63	0.63	0.62 ± 0.08
CM1 (D)	0.53	0.50	0.53	0.56	0.67	0.53	0.56	0.67	0.46	0.67	0.57 ± 0.07
CM2 (I-D-X)	0.67	0.63	0.43	0.31	0.33	0.59	0.46	0.71	0.63	0.71	0.54 ± 0.14
CM2 (I-X-D)	0.67	0.53	0.53	0.36	0.63	0.43	0.47	0.63	0.63	0.67	0.55 ± 0.10
CM2 (X-I-D)	0.67	0.71	0.57	0.36	0.31	0.40	0.62	0.56	0.70	0.44	0.53 ± 0.14
CM3	0.53	0.47	0.53	0.22	0.50	0.59	0.53	0.74	0.63	0.57	0.53 ± 0.12
CM4 [97]	0.53	0.43	0.47	0.22	0.36	0.44	0.14	0.63	0.50	0.59	0.43 ± 0.15
CM5 [45]	0.40	0.44	0.40	0.22	0.46	0.27	0.20	0.63	0.40	0.61	0.40 ± 0.14
CM6 [48]	0.63	0.62	0.62	0.67	0.62	**0.67**	0.50	0.75	0.62	0.63	0.63 ± 0.06
CM7 [116]	0.71	0.77	0.71	0.57	0.33	0.57	0.67	0.71	0.75	**0.78**	0.66 ± 0.13
CM8 [50]	0.50	0.50	0.50	0.50	0.40	0.50	0.50	0.72	0.61	0.61	0.53 ± 0.08

(a) True positive example

(b) True negative example

(c) False positive example

(d) False negative example

Figure 5.1.6: Some examples of estimation results for Par1. Figures (a) and (b) demonstrate that our method (I-X-D) accurately estimated the true emotion. Conversely, Figures (c) and (d) indicate instances where our method (I-X-D) incorrectly estimated the emotion.

5.1.3.2 Performance Evaluation

The results of the experiment are presented in Table 5.1.2. This table displays the F1-measure results, confirming that our method consistently outperforms all CMs on average. In this experiment, we explored various combinations of CNN features, denoted as "ours (a-b-c)". Here, a, b, and c stand for one of X (Xception), I (Inception-Resnet-v2), and D (Densenet201), and "ours (a-b-c)" represents the order of CNN features in the proposed method. As indicated in Table 5.1.2, our method surpasses the CMs in every combination. However, no significant difference was observed based on the order of CNN feature combination.

A comparative analysis between our approach and CM1 showcased the efficacy of incorporating temporal variations in brain activity and visual attention. Additionally, comparisons with CM2 and CM3 revealed the effectiveness of collaboratively using brain activity and gaze information. When compared with CM4, which utilizes traditional feature fusion [98] with other features extracted from gaze data [97], our approach proved to be more effective than CM4 in the multi-modal feature fusion framework. CM4 utilizes and fNIRS gaze features without incorporating temporal changes. This confirms the effectiveness of CCA-based feature fusion with GIT-based feature extraction for gaze data for estimating emotions. Besides, comparisons with

CMs 5-8 demonstrated that our approach is more valid than other frameworks. Although CMs 5 and 8 were methods estimating emotions by collaboratively using brain and gaze information, their outcomes were not satisfactory. This may be attributed to the limited training data for each category, which hindered the optimization of Deep CCA or LSTM, leading to insufficient training and lower estimation accuracy. On the contrary, the unsupervised learning method, BDAE, employed by CMs 6 and 7 required a smaller volume of training data in comparison to Deep CCA and LSTM, rendering it relatively optimized.

Figure 5.1.6 illustrates the estimation outcomes for a single participant. Figures 5.1.6 (a) and (b) showcase images where our method (I-X-D) accurately estimated emotions, while Fig. 5.1.6 (c) and (d) showcase images where our method (I-X-D) inaccurately estimated emotions. The images for which our method (I-X-D) accurately estimated emotions manifest distinctions in brightness and depicted objects. Notably, the image in Fig. 5.1.6 (a) is predominantly bright, with an object likely to evoke positive emotions in most individuals. Besides, the image in Fig. 5.1.6 (b) is predominantly dark, featuring an object that would likely elicit negative emotions in most individuals. Thus, these images are easily categorized as authentic representations of emotions. In contrast, Fig. 5.1.6 (c) features a predominantly dark background with an object resembling a flower. Typically, flowers are linked to positive emotions. Consequently, our method (I-X-D) might predict a positive emotion considering the attributes associated with a flower. The image depicted in Fig. 5.1.6 (d) is predominantly characterized by white and black colors and features a fox. Considering that the monochromatic arrangement might elicit negative emotions in individuals, our method (I-X-D) predicts a negative emotion when the participant observes the image in Fig. 5.1.6 (d).

Based on the qualitative and quantitative assessments described above, we can confirm the effectiveness of our approach in estimating human emotions and identify its limitations.

5.1.4 Conclusions

In this chapter, we introduced a human-centric emotion estimation method that maximizes correlation, taking into account temporal changes in both brain activity and visual attention. Our focus is on two types of biological data representing the temporal changes in visual atten-

tion towards objects in images and brain activity. To address these data, two feature extraction networks are constructed. Gaze-based features are transformed by maximizing correlations with fNIRS features, incorporating consideration of temporal changes through fourth mode of GIT. The primary contribution of this chapter lies in the CCA-based transformation of gaze-based CNN features and fNIRS features. Consequently, our approach accomplishes emotion estimation solely relying on gaze information, thereby eliminating the necessity for brain activity data during the test phase. The efficacy of human-centric emotion estimation has been substantiated through experimental findings.

Chapter 5.2

Human Emotion Recognition Using Multi-modal Biological Data Based on Time Lag-considered Correlation Maximization

5.2.1 Introduction

Chapter 5.1 presents the method for recognizing human emotions using multiple types of biological data. On the other hand, humans collect information through their eyes, and this information is subsequently processed in the brain. The visual stimuli perceived by the human eyes undergo transmission to the brain through neurotransmitters, leading to a time gap between gaze data and brain activity data [138], as depicted in Fig. 5.2.1. Previous studies, however, neglect such a time lag, merely combining features computed from brain activity and gaze data without considering the temporal misalignment. Consequently, these studies focus on integrating features not aligned with the same visual stimuli but rather from different stimuli between brain activity and gaze data, thereby constraining the expressive capacity of the features, which are calculated by integration methods in such studies, in capturing emotions recalled by humans. For addressing this gap, it is essential to devise an integration methodology that takes into account the temporal misalignment between brain activity and gaze data to comprehend the mechanism behind human emotions. Additionally, considering the heavy burden associated with acquiring brain activity data in daily life, we restrict the use of brain activity data solely to the training phase.

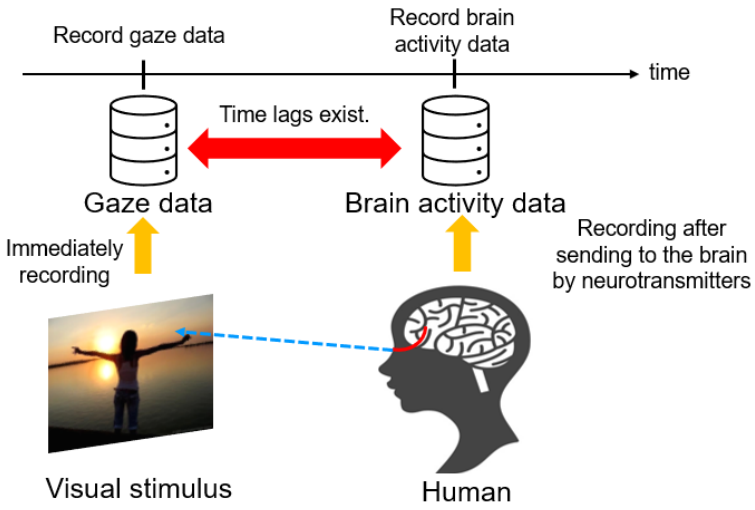

Figure 5.2.1: Concept figure of the time lag between gaze and brain activity data. The existence of a temporal delay arises from the transmission of visual stimuli, captured by human eyes, to the brain through neurotransmitters.

This chapter introduces a multi-modal human emotion recognition method grounded in time lag-considered correlation maximization. The time lag between gaze and brain activity data is contingent on the reaction time of the neuron, which may vary among individuals. Our proposed method aims to efficiently handle this time lag to model the reaction time of the neuron. The main emphasis in this chapter is the weighted correlation maximization to enhance feature integration, taking the time lag into consideration. More precisely, brain activity and gaze features are extracted, and transformation vectors are calculated to transform both features into a space shared by the these features. These transformation vectors are obtained by solving the maximization problem for correlations weighted by considering the displacement by time. To implement such correlation maximization considering the time lag, we extend the CCA framework. CCA is a simple transformation through the linear transformation, and our extension introduces a structure that explicitly recognizes the time lag, capturing the linear shift between brain activity and gaze. Within the CCA framework, latent features are obtained through the optimization of transformation vectors based on the correlation of multiple inputs. Our emphasis is on the correlation in the CCA framework, and time lag-considered weights are introduced into such correlation. In our approach, a specific distribution is assumed for generating time lag-considered weights. Once the transformation vectors are computed from the training data,

in the inference phase, the acquisition of brain activity data, which is burdensome, becomes unnecessary. Lastly, human emotions are recognized by utilizing features integrated through the transformation vectors. The distinctive contribution of this chapter lies in the novel construction of a human emotion recognition method that considers the time lag for making closer to a genuine understanding of the mechanism underlying the occurrence of human emotions.

5.2.2 Time Lag-considered Correlation Maximization for Human Emotion Recognition

This section describes the multi-modal human emotion recognition approach, which incorporates consideration for time lag in correlation maximization. Initially, we compute sequential features derived from gaze and brain activity data. For gaze features, we employ the GIT-based method, which is presented in Section 5.1.2.1, designed to capture visual information perceived by individuals. Additionally, we utilize functional near-infrared spectroscopy (fNIRS) for brain activity data, chosen for its superior temporal resolution compared to functional magnetic resonance imaging (fMRI). Notably, fNIRS data are considered more resilient to external activities such as eye blinks during image viewing, as opposed to electroencephalogram (EEG) data [124]. Various studies have explored the connection between fNIRS data and human emotions [126–128]. Hence, we employ fNIRS and gaze features as multi-modal features. It is crucial to emphasize that the proposed method exclusively utilizes fNIRS features for calculating the transformation vectors during the training phase. Subsequently, the gaze and fNIRS features undergo integration based on the time lag-considered Canonical Correlation Analysis (TlCCA), enabling the computation of latent features that encompass the commonalities between the two feature types. The efficacy of incorporating a time lag in the correlation-based integration of multimedia data and Twitter tweets has been demonstrated [139]. In this section, the TlCCA is mainly explained, while the feature extraction is explained in Sections 5.1.2.1 and 5.1.2.2.

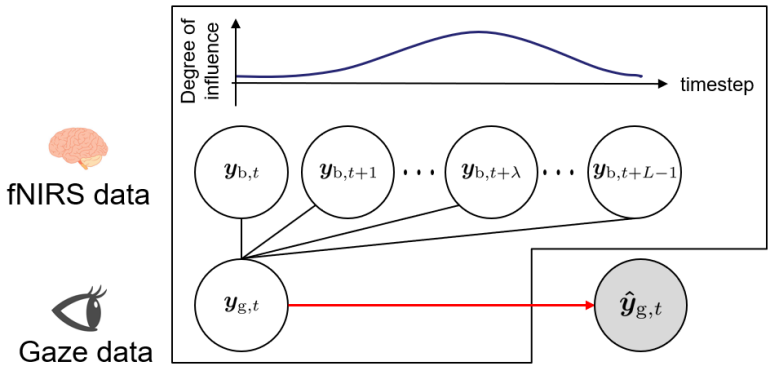

Figure 5.2.2: Outline of TlCCA. We incorporate weights that account for the time lag into the conventional correlation. In this figure, $\boldsymbol{y}_{\mathrm{g},t}$ and $\boldsymbol{y}_{\mathrm{b},t}$ represent $\boldsymbol{y}_{\mathrm{gaze},t}$ and $\boldsymbol{y}_{\mathrm{brain},t}$, respectively. Additionally, white and gray circles denote observed and unobserved variables. λ and L signify the peak and range of the time lag.

5.2.2.1 TlCCA-based Feature Integration

Figure 5.2.2 presents outline of TlCCA. We construct brain and gaze features as the following expression:

$$\boldsymbol{Y}_p = [\boldsymbol{y}_{p,1}, \boldsymbol{y}_{p,2}, \ldots, \boldsymbol{y}_{p,d_t}], \quad p = \{\mathrm{gaze}, \mathrm{brain}\}. \tag{5.2.1}$$

We make the assumption that fNIRS data are recorded a few seconds after gaze data are recorded for a given stimulus, introducing a time lag in the fNIRS data relative to the gaze data. Besides the time lag, the impact of visual stimuli may persist in the fNIRS data in the subsequent time step. In this context, we posit that human-obtained visual stimuli are promptly recorded in the gaze data, while the influence of these stimuli on fNIRS data follows a Poisson distribution. The TlCCA facilitates the computation of latent features from gaze and brain activity features while considering these assumptions about the time lag. Specifically, the optimization of the transformation vector set $\boldsymbol{w} = \{\boldsymbol{w}_{\mathrm{gaze}}, \boldsymbol{w}_{\mathrm{brain}}\} \in (\mathbb{R}^{d_{\mathrm{gaze}}}, \mathbb{R}^{d_{\mathrm{brain}}})$ is carried out using training feature sets $\{\boldsymbol{Y}_{\mathrm{gaze},n}, \boldsymbol{Y}_{\mathrm{brain},n}\}_{n=1}^{N}$ ($n = 1, 2, \ldots, N$; N representing the number of training data) as the

following equations:

$$\hat{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} \boldsymbol{w}_{\text{gaze}}^{\top} \sum_{n=1}^{N} \boldsymbol{C}_n^{\text{b}} \boldsymbol{w}_{\text{brain}} \tag{5.2.2}$$

$$\text{s.t. } \boldsymbol{w}_{\text{gaze}}^{\top} \boldsymbol{C}_n^{\text{gaze}} \boldsymbol{w}_{\text{gaze}} \boldsymbol{w}_{\text{brain}}^{\top} \boldsymbol{C}_n^{\text{brain}} \boldsymbol{w}_{\text{brain}} = 1, \forall n, \tag{5.2.3}$$

where $\boldsymbol{C}_n^{\text{gaze}}$ and $\boldsymbol{C}_n^{\text{brain}}$ represent the variance matrices of gaze and fNIRS features, respectively. The computation of these variance matrices is conducted as follows:

$$\boldsymbol{C}_n^p = \boldsymbol{Y}_{p,n} \boldsymbol{Y}_{p,n}^{\top}, \quad p = \{\text{gaze}, \text{brain}\}. \tag{5.2.4}$$

Moreover, $\boldsymbol{C}_n^{\text{b}}$ denotes the covariance matrix, accounting for the time lag between gaze and fNIRS features, and is computed as follows:

$$\boldsymbol{C}_n^{\text{b}} = \frac{1}{\sum_{l=0}^{L} e^{-\lambda} \lambda^l / l!} \sum_{l=0}^{L} \frac{e^{-\lambda} \lambda^l}{l!} \boldsymbol{Y}_{\text{gaze},n,l} \boldsymbol{Y}_{\text{brain},n,0}^{\top}, \tag{5.2.5}$$

where λ serves as a shape parameter for the Poisson distribution, determining the focal point of strongest influence of visual stimuli on the fNIRS features. L represents a hyperparameter dictating the number of timesteps over which visual stimuli continue to impact fNIRS features. It is important to note that the features are mean-normalized, and we construct the new feature set as following expression:

$$\boldsymbol{Y}_{p,n,l} = [\boldsymbol{y}_{p,n,L-l}, \boldsymbol{y}_{p,n,L+1-l}, \ldots, \boldsymbol{y}_{p,n,d_t-l}], \tag{5.2.6}$$

where $l = 0, 1, \ldots, L-1$. To solve Eq. (5.2.2), we employ the Lagrange multiplier method as follows:

$$F = \boldsymbol{w}_{\text{gaze}}^{\top} \sum_{n=1}^{N} \boldsymbol{C}_n^{\text{b}} \boldsymbol{w}_{\text{brain}} - \sum_{p} \beta_p \left(\sum_{n=1}^{N} \boldsymbol{w}_p^{\top} \boldsymbol{C}_n^p \boldsymbol{w}_p - 1 \right), \tag{5.2.7}$$

where $p = \{\text{gaze}, \text{brain}\}$, and β_{gaze} and β_{brain} represent the Lagrange coefficients. By computing $\partial F/\partial \boldsymbol{w}_{\text{gaze}} = 0$ and $\partial F/\partial \boldsymbol{w}_{\text{brain}} = 0$, we derive the eigenvalue problem as follows:

$$\begin{bmatrix} \boldsymbol{O} & \sum_{n=1}^{N} \boldsymbol{C}_n^{\text{b}\top} \\ \sum_{n=1}^{N} \boldsymbol{C}_n^{\text{b}} & \boldsymbol{O} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{\text{gaze}} \\ \boldsymbol{w}_{\text{brain}} \end{bmatrix} = \beta \begin{bmatrix} \sum_{n=1}^{N} \boldsymbol{C}_n^{\text{gaze}} & \boldsymbol{O} \\ \boldsymbol{O} & \sum_{n=1}^{N} \boldsymbol{C}_n^{\text{brain}} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_{\text{gaze}} \\ \boldsymbol{w}_{\text{brain}} \end{bmatrix}, \quad (5.2.8)$$

where $\beta = 2\beta_{\text{gaze}} = 2\beta_{\text{brain}}$. Solving Eq. (5.2.8) straightforwardly yields the optimal transformation vector pair $(\hat{\boldsymbol{w}}_{\text{brain}}, \hat{\boldsymbol{w}}_{\text{brain}})$. It is noteworthy that while several solution sets $(\beta, \hat{\boldsymbol{w}}_{\text{brain}}, \hat{\boldsymbol{w}}_{\text{brain}})$ may emerge, β signifies the efficiency of transformation vectors $(\hat{\boldsymbol{w}}_{\text{brain}}, \hat{\boldsymbol{w}}_{\text{brain}})$. Therefore, we arrange the eigenvalues β and select the top d_{latent} $(\leq \min(d_{\text{gaze}}, d_{\text{brain}}))$ eigenvalues along with their corresponding transformation vectors. The transformation matrices $\hat{\boldsymbol{W}}_{\text{gaze}} \in \mathbb{R}^{d_{\text{gaze}} \times d_{\text{latent}}}$ and $\hat{\boldsymbol{W}}_{\text{brain}} \in \mathbb{R}^{d_{\text{brain}} \times d_{\text{latent}}}$ are utilized for the actual transformation.

In the inference phase, where only gaze data is employed, the transformed features $\hat{\boldsymbol{Y}}_{\text{gaze}} \in \mathbb{R}^{d_{\text{latent}} \times d_t}$ are computed with the transformation matrix $\hat{\boldsymbol{W}}_{\text{gaze}}$ as the following equation:

$$\hat{\boldsymbol{Y}}_{\text{gaze}} = \hat{\boldsymbol{W}}_{\text{gaze}}^{\top} \boldsymbol{Y}_{\text{gaze}}. \quad (5.2.9)$$

The proposed approach enables obtaining the transformed features while considering the time lag, as presented in Eq. (5.2.5). Lastly, human emotions are recognized by generating d_{emotion}-dimensional one-hot vectors $\boldsymbol{e} \in \mathbb{R}^{d_{\text{emotion}}}$, where each element corresponds to a specific human emotion as

$$\boldsymbol{e} = \boldsymbol{f}(\hat{\boldsymbol{Y}}_{\text{gaze}}), \quad (5.2.10)$$

where $\boldsymbol{f}(\cdot)$ is the classifier trained with the transformed features derived from the training data.

5.2.3 Experiments

This experiment adopted the same dataset as Chapter 5.1, and the details are presented in Section 5.1.3.1

We employed seven comparative methods for evaluation. As detailed in Table 5.2.1, two of them, referred to as Ablations 1 and 2, focused on either gaze or fNIRS features, constituting

Table 5.2.1: Characteristics of our approach and comparative methods.

	Features		Time	
	Gaze	fNIRS	Change	Lag
Ablation 1		✓	✓	
Ablation 2	✓		✓	
Deep CCA [45]	✓	✓		
BDAE [48]	✓	✓		
BLSTM [50]	✓	✓	✓	
MVAE [140]	✓	✓	✓	
CCA with GIT [114]	✓	✓	✓	
Our method	✓	✓	✓	✓

uni-modal approaches. It is worth noting that these two methods applied principal component analysis [133] to reduce feature dimensions to d_{latent}. The remaining five methods were drawn from previous studies [45, 48, 50, 114, 140], utilizing various feature integration techniques such as Deep CCA [46], BDAE [49], BLSTM [51], CCA with GIT [114], and multi-view variational autoencoder (MVAE) [141] to integrate multi-modal features. Notably, [45, 48] lacked mechanisms to consider both time lag and time changes, computing each feature when $d_t = 1$ in Section 5.1.2.1. Additionally, in the inference phase of [45, 48, 50, 140], both gaze and brain activity features were obligatory, while in [114] and our proposed method, only gaze data were required. Support Vector Machine (SVM) [105] was consistently employed as the final classifier $\boldsymbol{f}(\cdot)$ in each method. The hyperparameters L, λ, d_{brain}, d_{gaze}, and d_{latent} were set to 5, 1, 440, 2048, and 50, respectively. To conduct the performance evaluation, we considered Recall, Precision, F1-score, and Accuracy metrics.

Table 5.2.2 presents the mean results of each method. Through a comparison of our method with Ablations 1 and 2, we validate the efficacy of jointly utilizing gaze and fNIRS features for human emotion recognition. While Ablation 2 excels in the "Recall" evaluation metric, the proposed method outperforms others in other evaluation metrics. Notably, despite Ablation 2 having a lower Precision than the proposed method, the recognition ability of the proposed method is superior, as evidenced by a higher F1-score, which is the harmonic mean of Recall and Precision. Furthermore, when contrasting our method with [45, 48], the effectiveness of

Table 5.2.2: Mean results of each method.

	Recall	Precision	F1-score	Accuracy
Ablation 1	0.30	0.48	0.65	0.52
Ablation 2	**0.83**	0.74	0.76	0.77
Deep CCA [45]	0.57	0.54	0.53	0.58
BDAE [48]	0.29	0.64	0.55	0.57
BLSTM [50]	0.37	0.31	0.44	0.44
MVAE [140]	0.49	0.55	0.52	0.57
CCA with GIT [114]	0.63	**0.85**	0.67	0.74
Our method	0.75	0.84	**0.78**	**0.81**

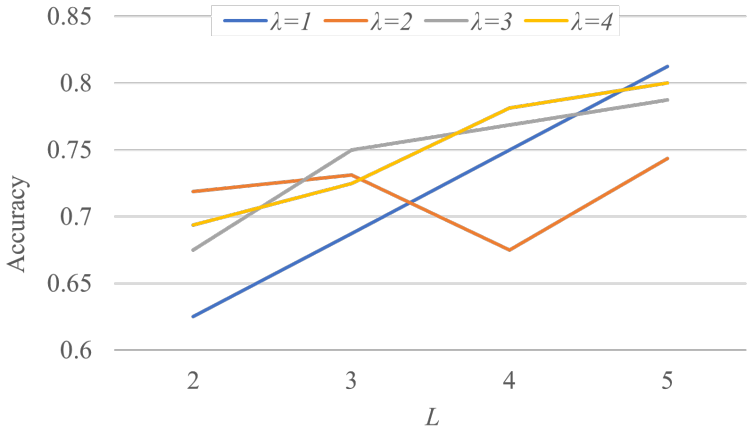

Figure 5.2.3: Variations in the mean accuracy of the proposed approach regarding λ and L.

considering temporal changes becomes apparent. In comparison with [114], where CCA with GIT performs well in the "Precision" evaluation metric, effectiveness of the proposed method is evident in the same context as Recall of Ablation 1. Lastly, the comparison with [140] reaffirms the effectiveness of considering the time lag between gaze and fNIRS data, the primary focus of this chapter. In the "Recall" evaluation metric, Ablation 2 outperforms other methods, but the proposed method surpasses others in alternative evaluation metrics.

Figure 5.2.3 illustrates the variations in mean accuracy for the proposed method concerning λ and L. Notably, for any λ, the accuracy is optimal at $L = 5$. Specifically, the highest accuracy is achieved when $\lambda = 1$, signifying that the peak of the time lag is to one timestep, that is one second. Our results align closely with results reported in the other studies [138, 142] within the

field of brain computing. Consequently, the proposed method effectively captures the human cognition process by incorporating the time lag.

5.2.4 Conclusions

This chapter presented the multi-modal method for recognizing human emotion through the TlCCA-based feature integration, which accounts for the time lag between fNRIS and gaze data. Specifically, we incorporated the mechanism to address the time lag-considered weights into correlation of the CCA scheme, by assuming that the impact of visual stimuli on fNIRS data adheres to the Poisson distribution. Through the adoption of TlCCA, we have innovatively constructed the time lag-considered human emotion recognition method.

Chapter 5.3

Multi-view Variational Recurrent Neural Network for Human Emotion Recognition Using Multi-modal Biological data

5.3.1 Introduction

Biological data measurements frequently contain inherent noises, including measurement errors, which can perturb methods based on deterministic machine learning and result in misinterpretation of the relationships between human emotions and biological data. In the domain of brain-machine interaction, it is common to utilize probabilistic machine learning to mitigate the impact of noise in brain activity data, leading to successful outcomes [143–145]. Hence, the utilization of probabilistic machine learning for integration is expected to yield stable integrated features, ensuring resilient human emotion recognition by accounting for the impact of noises. Therefore, the integration method needs to jointly consider three key characteristics of biological data: 1) the relationship between explicit and implicit information such as brain activity and gaze, 2) temporal changes associated with emotions recalled by humans, and 3) the potential impact of noises.

This chapter introduces the Multi-view Variational Recurrent Neural Network (MvVRNN) for multi-modal human emotion recognition. In the proposed approach, diverse features computed from different biological data are integrated using MvVRNN. Subsequently, a lightweight classifier is optimized using shared latent features obtained by MvVRNN to realize multi-modal

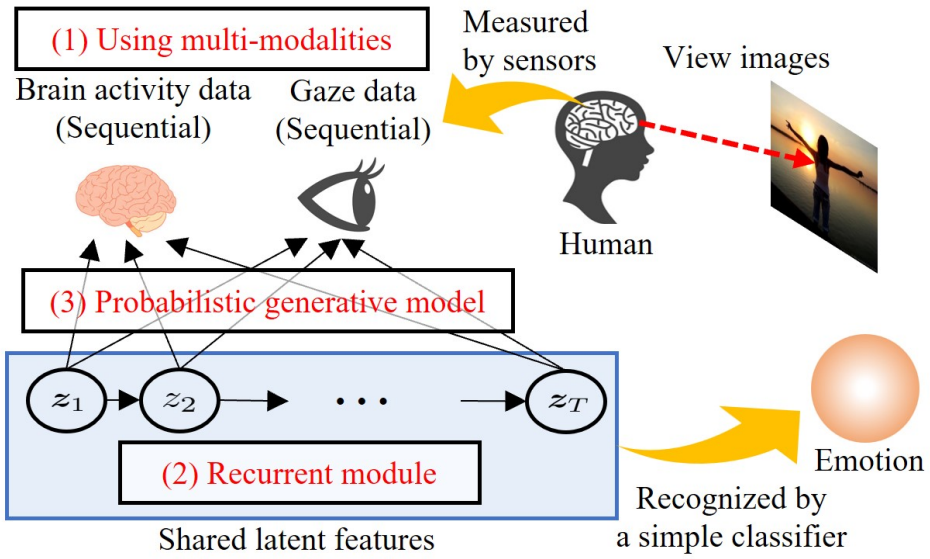

Figure 5.3.1: Context and emphasis in this chapter. Within our approach, we introduce the MvVRNN specifically for human emotion recognition when individuals view images. This is achieved by emphasizing 1) the utilization of multiple types of biological data, 2) the incorporation of the recurrent module designed for sequential data, and 3) the integration of the probabilistic generative model.

human emotion recognition. MvVRNN, which is a generative model, incorporates mechanisms including 1) the integration of multi-modal information encompassing implicit and explicit human states, 2) a recurrent module for sequential data, and 3) variational approximation based on the Gaussian distribution, jointly addressing the three aforementioned characteristics of biological data as illustrated in Fig. 5.3.1. In MvVRNN, we posit that shared latent features generate multi-modal sequential data for integration with temporal dependencies and probabilistic variation. It is important to note distinctions between the proposed MvVRNN and the multi-view variational autoencoder with the recurrent module [146]. For instance, our model conditions recurrent modules on previous latent variables to consider temporal dependencies of them. The main contributions of this chapter are summarized as follows:

- By applying MvVRNN, which is newly derived for multi-modal human emotion recognition, to multi-modal sequential data, dependencies of latent features across timesteps and relationships between multiple views can be considered while reducing the impact of noises through variational approximation.

- Integrated features from MvVRNN offer enhanced representational power for human emotions via the adoption of distributed latent features.
- It is indicated the efficiency of MvVRNN for recognizing human emotions through comparisons with other feature integration methods.

(a) Calculation of the prior distribution (b) Generation process of input data (c) Recurrence (d) Calculation of the posterior distribution

Figure 5.3.2: Graphical representations of the MvVRNN are presented. Specifically, (a) depicts the process of calculating the prior distribution for the shared latent features presented in Section 5.3.2.1, (b) depicts the generation process of multi-modal sequential data by decoding the shared latent features presented in Section 5.3.2.2, (c) depicts the recurrent component presented in Section 5.3.2.3, and (d) depicts the calculation of the posterior distribution for feature integration presented in Section 5.3.2.4.

5.3.2 Emotion Recognition Using MvVRNN

This section delineates the proposed method for emotion recognition utilizing the MvVRNN-based feature integration. Initially, our approach combines various features from distinct biological data sources while considering their temporal dynamics and variability. Subsequently, we employ a lightweight classifier to facilitate the recognition of human emotions. The input data from the mth modality ($m = 1, 2, \ldots, M$; M representing the number of modalities) is denoted as $\boldsymbol{x}_t^{(m)} \in \mathbb{R}^{d_m}$. It is crucial to note that the class label y associated with user feedback is present in the data. However, this is utilized not during the training of MvVRNN but for training a separate classifier. In the following, we provide a detailed explanation of the MvVRNN. The graphical representations of the MvVRNN are illustrated in Fig. 5.3.2.

5.3.2.1 Prior Distribution

In MvVRNN, it is posited that the multi-modal sequential input data arise from shared latent features at each timestep. Additionally, these shared latent features are considered independent of one another and are assumed to conform to a Gaussian distribution for ease of mathematical treatment. The prior distribution of the shared latent features $\boldsymbol{z}_t \in \mathbb{R}^{d_z}$ at timestep t is computed as follows:

$$\boldsymbol{z}_t \sim \mathcal{N}(\boldsymbol{\mu}_{\text{pri},t}, \text{diag}(\boldsymbol{\sigma}^2_{\text{pri},t})), \tag{5.3.1}$$

$$[\boldsymbol{\mu}_{\text{pri},t}, \boldsymbol{\sigma}_{\text{pri},t}] = f(\boldsymbol{h}_{t-1}^{(1)}, \boldsymbol{h}_{t-1}^{(2)}, \ldots, \boldsymbol{h}_{t-1}^{(M)}), \tag{5.3.2}$$

where $\boldsymbol{\mu}_{\text{pri},t}$ and $\boldsymbol{\sigma}_{\text{pri},t}$ represent the parameters of the prior distribution, and their calculation is performed by the function f. The $\boldsymbol{h}_{t-1}^{(m)}$ variables denote the previous state variables of the recurrent module corresponding to the mth modality. In MvVRNN, we condition the prior distribution on these previous state variables, following the approach of the Variational Recurrent Neural Network (VRNN) [147], and take into account the dependencies of the shared latent features at neighboring timesteps.

5.3.2.2 Probabilistic Generation Process

MvVRNN produces input data at each timestep by decoding the shared latent features. More precisely, in MvVRNN, the generation of multi-modal input data involves preparing decoders $\{\text{Dec}^{(m)}\}_{m=1}^{M}$ for each modality. The generation processes, conditioned on the shared latent features z_t and the previous state variables of the recurrent module $\boldsymbol{h}_{t-1}^{(m)}$, are expressed by the following equations:

$$\boldsymbol{x}_t^{(m)}|z_t \sim \mathcal{N}(\boldsymbol{\mu}_t^{(m)}, \text{diag}(\boldsymbol{\sigma}_t^{(m)^2})), \tag{5.3.3}$$

$$[\boldsymbol{\mu}_t^{(m)}, \boldsymbol{\sigma}_t^{(m)}] = \text{Dec}^{(m)}(g_z(z_t), \boldsymbol{h}_{t-1}^{(m)}), \tag{5.3.4}$$

where $\boldsymbol{\mu}_t^{(m)}$ and $\boldsymbol{\sigma}_t^{(m)}$ denote the parameters of the generated distribution. Additionally, g_z represents the function that transforms the shared latent features z_t. While the shared latent features are common across all modalities, we generate input data specific to each modality by establishing dedicated decoders for each modality. The integration of input data is achieved by treating the shared latent features as integrated features, taking into account their temporal dependencies.

5.3.2.3 Recurrent Module

To capture the temporal dynamics of multi-modal sequential input data, we incorporate a recurrent module, such as the Long Short-term Memory (LSTM) [51] and the Gated Recurrent Unit (GRU) [148]. Concretely, the multi-modal sequential input data are fed into the respective recurrent module for each modality, and the process is outlined as follows:

$$\boldsymbol{h}_t^{(m)} = g_h(g_{x^{(m)}}(\boldsymbol{x}_t^{(m)}), g_z(z_t), \boldsymbol{h}_{t-1}^{(m)}), \tag{5.3.5}$$

where g_h represents the nonlinear mapping function, and $g_{x^{(m)}}$ is the function responsible for transforming the input data of the mth modality. In the proposed MvVRNN, a dedicated recurrent module is designed for each modality to effectively capture the temporal dynamics inherent in each modality. Consequently, the resultant state variables $\boldsymbol{h}_t^{(m)}$ encompass temporal information derived from each input data, facilitating the extraction of sequential data characteristics.

The incorporation of the recurrent module addresses the specific focus (2) presented in Section 5.3.1.

5.3.2.4 Posterior Distribution

The purpose of MvVRNN in this chapter is to effectively integrate multi-modal input data while preserving temporal information. Therefore, the proposed method needs to derive shared latent features conditioned on input data, specifically, the posterior distribution of the shared latent features. Calculating the posterior distribution in the Variational AutoEncoder (VAE) [149] framework may not be feasible analytically, and an approximate distribution is typically employed. Similarly, we utilize an approximate distribution for the posterior distribution and train the model by minimizing the Kullback-Leibler divergence to converge towards the true posterior distribution. The approximate distribution is expressed as follows:

$$z_t|\boldsymbol{x}_t^{(1)}, \boldsymbol{x}_t^{(2)}, \ldots, \boldsymbol{x}_t^{(M)} \sim \mathcal{N}(\boldsymbol{\mu}_{\text{post},t}, \text{diag}(\boldsymbol{\sigma}_{\text{post},t}^2)), \tag{5.3.6}$$

$$[\boldsymbol{\mu}_{\text{post},t}, \boldsymbol{\sigma}_{\text{post},t}] = \text{Enc}\Big(g_{x^{(1)}}(\boldsymbol{x}_t^{(1)}), g_{x^{(2)}}(\boldsymbol{x}_t^{(2)}), \ldots, g_{x^{(M)}}(\boldsymbol{x}_t^{(M)}), \boldsymbol{h}_{t-1}^{(1)}, \boldsymbol{h}_{t-1}^{(2)}, \ldots, \boldsymbol{h}_{t-1}^{(M)}\Big), \tag{5.3.7}$$

where $\boldsymbol{\mu}_{\text{post},t}$ and $\boldsymbol{\sigma}_{\text{post},t}$ represent the parameters of the approximate distribution. Enc$(\cdot)$ denotes the encoder, typically implemented as a neural network. Lastly, the shared latent features $\boldsymbol{z}_{t=1}^{T}$, encoded from multi-modal sequential input data, serve as integrated features. The calculation of the posterior distribution for the shared latent features allows us to achieve our objectives outlined in Section 5.3.1, specifically, focuses (1) and (3).

5.3.2.5 Objective Function

The aforementioned model is optimized using variational inference techniques. Specifically, we maximize the marginal likelihood based on the variational lower bound. The variational lower bound in MvVRNN necessitates the joint distribution and the approximate distribution across timesteps, denoted as $p(\boldsymbol{x}_{\leq T}^{(1)}, \boldsymbol{x}_{\leq T}^{(2)}, \ldots, \boldsymbol{x}_{\leq T}^{(M)}, \boldsymbol{z}_{\leq T})$ and $q(\boldsymbol{z}_{\leq T}|\boldsymbol{x}_{\leq T}^{(1)}, \boldsymbol{x}_{\leq T}^{(2)}, \ldots, \boldsymbol{x}_{\leq T}^{(M)})$, respectively, to calculate the variational lower bound, similar to the approach in VAE. It is important to note that $\boldsymbol{x}_{\leq t} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$. Leveraging Equations (5.3.1), (5.3.3), and (5.3.5), we compute

the joint distribution conditioned on the state variables of the recurrent module as follows:

$$p(\boldsymbol{x}_{\leq T}^{(1)}, \boldsymbol{x}_{\leq T}^{(2)}, \ldots, \boldsymbol{x}_{\leq T}^{(M)}, z_{\leq T}) = \prod_{t=1}^{T} p(z_t|z_{<t}, \boldsymbol{x}_{<t}^{(1)}, \boldsymbol{x}_{<t}^{(2)}, \ldots, \boldsymbol{x}_{<t}^{(M)}) \times \prod_{m=1}^{M} p(\boldsymbol{x}_t^{(m)}|z_{\leq t}, \boldsymbol{x}_{<t}^{(1)}, \boldsymbol{x}_{<t}^{(2)}, \ldots, \boldsymbol{x}_{<t}^{(M)}). \quad (5.3.8)$$

Furthermore, utilizing Equations (5.3.3) and (5.3.6), we compute the approximate distribution conditioned on the state variables of the recurrent module as follows:

$$q(z_{\leq T}|\boldsymbol{x}_{\leq T}^{(1)}, \boldsymbol{x}_{\leq T}^{(2)}, \ldots, \boldsymbol{x}_{\leq T}^{(M)}) = \prod_{t=1}^{T} q(z_t|z_{<t}, \boldsymbol{x}_{\leq t}^{(1)}, \boldsymbol{x}_{\leq t}^{(2)}, \ldots, \boldsymbol{x}_{\leq t}^{(M)}). \quad (5.3.9)$$

Equations (5.3.8) and (5.3.9) can be interpreted as factorization [147]. The variational lower bound, serving as the objective function, is defined by leveraging these relationships, as follows:

$$\mathbb{E}_q\Big[\sum_{t=1}^{T}\Big(\sum_{m=1}^{M} \log p(\boldsymbol{x}_t^{(m)}|\boldsymbol{x}_{<t}^{(1)}, \boldsymbol{x}_{<t}^{(2)}, \ldots, \boldsymbol{x}_{<t}^{(M)}, z_{\leq t}) - \mathbb{D}_{KL}(q(z_t|\boldsymbol{x}_{\leq t}^{(1)}, \boldsymbol{x}_{\leq t}^{(2)}, \ldots, \boldsymbol{x}_{\leq t}^{(M)}, z_{\leq t})\|p(z_t|\boldsymbol{x}_{<t}^{(1)}, \boldsymbol{x}_{<t}^{(2)}, \ldots, \boldsymbol{x}_{<t}^{(M)}, z_{<t}))\Big)\Big], \quad (5.3.10)$$

where $\mathbb{D}_{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. This objective function serves to jointly optimize the encoder, the decoder, and the recurrent module. The stochastic backpropagation technique is subsequently employed to train the parameters of MvVRNN. Specifically, the reparameterization trick [149] is utilized for optimizing the parameters included in distributions, following a similar approach to VAE. Thus, we concurrently optimize the trainable parameters across various functions of MvVRNN.

5.3.3 Experiments

This experiment adopted the same dataset as Chapter 5.1, and the details are presented in Section 5.1.3.1

To conduct comparative experiments, seven methods were implemented as outlined in Table 5.3.1. As ablation studies, we adopted VRNN-based approaches using only fNRIS or gaze-based features (VRNN-brain or VRNN-gaze). The five remaining methods were previously introduced for the multi-modal recognition of human emotions [45, 48, 50, 114, 140], with a

Table 5.3.1: Characteristics of each method.

	Multi-modality	Recurrence	Variational
VRNN-gaze		✓	✓
VRNN-brain		✓	✓
TC-MVAE [140]	✓		✓
TC-CCA [114]	✓		
Deep CCA [45]	✓		
BDAE [48]	✓		
BLSTM [50]	✓	✓	
MvVRNN	✓	✓	✓

Table 5.3.2: Evaluation results for each method.

	Recall	Precision	F1-score	Accuracy
VRNN-gaze	0.29	0.28	0.25	0.49
VRNN-brain	0.58	0.45	0.49	0.54
TC-MVAE [140]	0.49	0.55	0.52	0.57
TC-CCA [114]	0.63	**0.85**	0.67	0.74
Deep CCA [45]	0.57	0.54	0.53	0.58
BDAE [48]	0.29	0.64	0.55	0.57
BLSTM [50]	0.37	0.31	0.44	0.44
MvVRNN	**0.77**	0.67	**0.70**	**0.75**

primary focus on integrating brain activity and gaze-based features. We directly input these integrated features into Support Vector Machine (SVM) [105] for recognizing human emotions. Notably, human emotion recognition methods based on Deep CCA and BDAE [45, 48] lacked consideration for temporal changes in sequential data, necessitating the averaging of each feature across timesteps for these methods. The MvVRNN was optimized using ADAM [150], with the mini-batch size, learning rate, and epoch set to 8, 1.0×10^{-5}, and 420, respectively. Additionally, the dimension of the shared latent features was configured as 16. Evaluation metrics such as "Recall", "Precision", "F1-score", and "Accuracy" were employed to assess the proposed and compared methods.

The experimental results presented in Table 5.3.2 compare the proposed MvVRNN with the other methods. Firstly, the comparison with VRNN-brain and VRNN-gaze confirms the superior effectiveness of multi-modality, incorporating both gaze and brain activity data, over

uni-modality in human emotion recognition. Besides, comparing the MvVRNN with BDAE, Deep CCA, and TC-CCA reveals that the evaluation results of the MvVRNN are the highest among these methods, substantiating the efficacy of the recurrence and variational mechanisms. Meanwhile, the recurrent module demonstrates utility in capturing temporal information in biological data when compared to the time-considered multi-modal variational autoencoder (TC-MVAE) [140]. Additionally, the variational approximation proves suitable for modeling uncertainty in biological data when contrasted with BLSTM. In this manner, both the recurrent module and variational approximation are shown to be effective, underscoring the suitability of their collaborative use in multi-modal human emotion recognition. Consequently, the efficacy of the MvVRNN for feature integration is validated in the context of human emotion recognition.

5.3.4 Conclusions

In this chapter, we introduce MvVRNN that is a novel model designed for recognizing human emotions using multiple tyes of biological gata. Specifically, MvVRNN is innovatively developed to integrate multi-modal sequential data with taking into account dependencies of latent variables across timesteps and relationships of multiple views and mitigating the impact of noises. Experimental results demonstrate the effectiveness of the newly introduced MvVRNN for feature integration in the context of human emotion recognition.

Chapter 6

Conclusions

As conclusions of this thesis, this chapter summarizes the proposition and clarifies future directions.

6.1 Summary of the Proposition

This section provides a summary of the proposition in this thesis. This thesis deals with the construction of machine learning models specific to personalized prediction of human perception toward visual stimuli. Biological data are used as the person-specific information, and several machine learning models suitable for biological data are presented.

In Chapter 3, three methods for the prediction of the personalized saliency map (PSM) with a limited amount of training data are presented. These methods focus on the similarities of visual attention between persons for the prediction of personalized salient regions in images from the limited amount of training data. To calculate such similarities, the images that persons commonly gazed at are needed. Hence, the adaptive image selection module considering object and visual attention is proposed and introduced into the PSM prediction model in a simple manner. As the PSM prediction models based on the similarities of visual attention between persons, the following three models are presented: 1) the weighted average-based model, 2) the Gaussian process regression-based model, and 3) the model using object-based similarities of gaze tendency. These methods steadily achieve improvement in performance.

In Chapter 4, two methods for gaze-based emotional category classification of images are

presented. For simultaneously analyzing the content of visual stimuli and human visual attention, the novel uniform representation including visual contents and gaze data is proposed. The constructed representations are the fourth-order tensors, and the machine learning-based tensor analysis is applied to them. By confirming the performance of emotion label estimation, such representation is indicated to contain both the visual contents and gaze data. As the gaze-based emotion label classification, CNN features are extracted from the constructed representation. CNN features are outputs of an intermediate layer of the pre-trained CNN and are well-known to their high representation ability. However, they do not necessarily have the high discrimination ability for our target domain, and multiple CNN features, which are extracted from multiple CNN models, are used. Experimental results show the effectiveness of these approaches.

In Chapter 5, three multi-modal methods for recognizing human emotions using several types of biological information are presented. Multiple types of biological data are used to compensate for information that is missing from a single type of biological data. Each type of biological data represents a different aspect of the human response, and the human perception can be more precisely predicted by collaboratively using them than one of them alone. To deal with the several types of biological data, feature integration methods are presented since biological data are pre-processed for calculating features suitable for each type of data before inputting machine learning models, generally. The first method simply focuses on the correlation-based feature integration treating several types of biological information. By using the canonical correlation analysis, heterogeneous features are transformed into the common feature spaces with properties of multiple input features. Transformed features are input to the simple machine learning model for predicting human perception. Besides, humans collect information through their eyes, and this information is subsequently processed in the brain. The visual stimuli perceived by the human eyes undergo transmission to the brain through neurotransmitters, leading to a time gap between gaze data and brain activity data Hence, the second method integrates features with considering the temporal misalignment of multiple biological data. Finally, the third method focuses on the other characteristics of biological signals. Specifically, this method realizes feature integration with considering the following three characteristics: 1) the relationship between explicit and implicit information such as brain activity and gaze, 2) temporal changes associated with emotions recalled by humans, and 3) the potential impact of noises. For simultaneously consid-

ering them, the multi-view variational recurrent neural network is newly derived. Experiments on datasets derived from personally acquired raw data showed the progressive improvement in performance.

The contributions of this thesis are the proposals of the several machine learning models specific to the biological data for predicting the human perception toward visual stimuli. The methods incorporates the mechanisms that can deal with the unique properties of biological data, and their effectiveness has been validated by conducting experiments on on datasets derived from personally acquired raw data and openly available datasets.

6.2 Future Directions

This section clarifies the future directions of this study.

Although the gaze and brain activity data are used as biological data in this thesis, there are a variety of biological data such as Electrocardiogram (ECG) and facial expression. With the advancements in sensor technologies, the sensors measuring biological data are becoming smaller and more inexpensive, and they are being used beyond the scope of research. Under these circumstances, there is a need to construct a unified model that can comprehensively handle biological data regardless of the format or type of data since individuals possess different sensors and have access to varying types of data. The establishment of such a unified model has the potential to address or accommodate issues pertaining to individual differences, which are currently processed in isolation in this thesis.

One of the purposes of the personalized prediction of human perception is to introduce person-specific information into various tasks such as recommendation and information retrieval. This thesis covers the construction of machine learning models specific to biological data, but their application to such tasks is beyond the scope. Hence, one of the future directions is the actual implementation of personalized prediction of human perception for real-world applications.

As mentioned in Chapter 1, the human responses depend on the kind of stimuli, and the contents of stimuli are important for personalized prediction of human perception. While, in the fields of computer vision and natural language processing, the very large-scale models, which can solve several tasks in a single model, have been constructed [151, 152]. These models

have successfully achieved advanced semantic understandings from images and text descriptions. Therefore, the utilization of such large-scale models may enhance the personalized prediction of human perception presented in this thesis.

To mitigate the substantial volume of data acquired from individuals, this thesis primarily explores the data similarities between individuals. In contrast, federated learning [153, 154] has been extensively researched for handling data specific to each individual. Integrating the principles of federated learning into the method presented in this thesis holds the promise of enhancing the efficiency of data utilization. However, when incorporating data from external sources, particularly from other individuals, significant privacy concerns may arise. Therefore, it is imperative to construct models that preserve privacy when utilizing data obtained from different individuals.

Bibliography

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[2] M. Yutaka, L. Yann, S. Maneesh, P. Doina, S. David, S. Masashi, U. Eiji, and M. Jun, "Deep learning, reinforcement learning, and world models," *Neural Networks*, vol. 152, pp. 267–275, 2022.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[5] A. C. Tsoi, "Recurrent neural network architectures: An overview," *International School on Neural Networks, Initiated*, pp. 1–26, 1997.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in *Proc. the Int'l Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.

[7] G. Yenduri, G. Srivastava, P. K. R. Maddikunta, R. H. Jhaveri, W. Wang, A. V. Vasilakos, T. R. Gadekallu, et al., "Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions," *arXiv preprint arXiv:2305.10435*, 2023.

[8] N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, H. Lu, R. Song, X. Gao, T. Xiang, et al., "Towards artificial general intelligence via a multimodal foundation model," *Nature Communications*, vol. 13, no. 1, pp. 3094, 2022.

[9] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, et al., "A comprehensive survey on pretrained foundation models: A history from bert to chatgpt," *arXiv preprint arXiv:2302.09419*, 2023.

[10] J. Liu, C. Yang, Z. Lu, J. Chen, Y. Li, M. Zhang, T. Bai, Y. Fang, L. Sun, P. S. Yu, et al., "Towards graph foundation models: A survey and beyond," *arXiv preprint arXiv:2310.11829*, 2023.

[11] D. Wang and X. Zhao, "Affective video recommender systems: A survey," *Frontiers in Neuroscience*, vol. 16, pp. 984404, 2022.

[12] Q. Liu, J. Hu, Y. Xiao, J. Gao, and X. Zhao, "Multimodal recommender systems: A survey," *arXiv preprint arXiv:2302.03883*, 2023.

[13] C. Vinola and K. Vimaladevi, "A survey on human emotion recognition approaches, databases and applications," *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, vol. 14, no. 2, pp. 24–44, 2015.

[14] E. Rusnandi, E. Winarko, and S. Azhari, "A survey on multimodal information retrieval approach," in *Proc. IEEE Int'l Conf. Smart Technology and Applications (ICoSTA)*, 2020, pp. 1–6.

[15] Y. Zhu, H. Yuan, S. Wang, J. Liu, W. Liu, C. Deng, Z. Dou, and J.-R. Wen, "Large language models for information retrieval: A survey," *arXiv preprint arXiv:2308.07107*, 2023.

[16] K. Sugata, T. Ogawa, and M. Haseyama, "Selection of significant brain regions based on MvGTDA and TS-DLF for emotion estimation," *IEEE Access*, vol. 6, pp. 32481–32492, 2018.

[17] H. Yoon and S. Chung, "EEG-based emotion estimation using Bayesian weighted-log-posterior function and perceptron convergence algorithm," *Computers in Biology and Medicine*, vol. 43, no. 12, pp. 2230–2237, 2013.

[18] K. Tai and T. Chau, "Single-trial classification of NIRS signals during emotional induction tasks: Towards a corporeal machine interface," *Journal of Neuroengineering and Rehabilitation*, vol. 6, no. 39, pp. 1–14, 2009.

[19] Y. Sasaka, T. Ogawa, and M. Haseyama, "Multimodal interest level estimation via variational Bayesian mixture of robust CCA," in *Proc. ACM Int'l Conf. Multimedia (ACM MM)*, 2016, pp. 387–391.

[20] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, vol. 20, no. 11, pp. 1254–1259, 1998.

[21] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[22] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 19, pp. 545–552, 2006.

[23] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, vol. 34, no. 1, pp. 194–201, 2012.

[24] Q. Peng, Y.-m. Cheung, X. You, and Y. Y. Tang, "A hybrid of local and global saliencies for detecting image salient region and appearance," *IEEE Trans. Systems, Man, and Cybernetics: Systems*, vol. 47, no. 1, pp. 86–97, 2016.

[25] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2014, pp. 2798–2805.

[26] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2014.

[27] M. Kümmerer, T. S. Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[28] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE/CVF Int'l Conf. Computer Vision (IEEE/CVF ICCV)*, 2015, pp. 262–270.

[29] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. G.-i. Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 27, pp. 2672–2680, 2014.

[31] X. Zhao, H. Lin, P. Guo, D. Saupe, and H. Liu, "Deep learning vs. traditional algorithms for saliency prediction of distorted images," in *Proc. IEEE Int'l Conf. Image Processing (IEEE ICIP)*, 2020, pp. 156–160.

[32] Y. Xu, N. Li, J. Wu, J. Yu, and S. Gao, "Beyond universal saliency: Personalized saliency prediction with multi-task CNN," in *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI)*, 2017, pp. 3887–3893.

[33] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Trans. Image Processing (IEEE TIP)*, vol. 27, no. 2, pp. 964–975, 2017.

[34] Y. Xu, S. Gao, J. Wu, N. Li, and J. Yu, "Personalized saliency and its prediction," *IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, vol. 41, no. 12, pp. 2975–2989, 2018.

[35] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, vol. 41, no. 3, pp. 740–757, 2018.

[36] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14:4, pp. 1–17, 2007.

[37] M. Kümmerer, T. Wallis, and M. Bethge, "How close are we to understanding image-based saliency?," *arXiv preprint arXiv:1409.7686*, 2014.

[38] M. Kümmerer, T. S. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16054–16059, 2015.

[39] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. ACM Int'l Conf. Multimedia (ACM MM)*, 2010, pp. 83–92.

[40] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proc. AAAI Conf. Artificial Intelligence*, 2015, vol. 29, pp. 381–388.

[41] V. Campos, B. Jou, and X. G.-i. Nieto, "From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction," *Image and Vision Computing*, vol. 65, pp. 15–22, 2017.

[42] S. Zhao, G. Ding, Y. Gao, X. Zhao, Y. Tang, J. Han, H. Yao, and Q. Huang, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Trans. Affective Computing (IEEE TAF)*, vol. 11, no. 4, pp. 574–587, 2018.

[43] S. Zhao, G. Ding, Y. Gao, and J. Han, "Approximating discrete probability distribution of image emotions by multi-modal features fusion," in *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI)*, 2017, pp. 4669–4675.

[44] S. Lee, C. Ryu, and E. Park, "OSANet: Object semantic attention network for visual sentiment analysis," *IEEE Trans. Multimedia (TMM)*, vol. 25, pp. 7139–7148, 2023.

[45] J. Qiu, W. Liu, and B. Lu, "Multi-view emotion recognition using deep canonical crrelation analysis," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2018, pp. 221–231.

[46] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int'l Conf. Machine Learning (ICML)*, 2013, pp. 1247–1255.

[47] K. Pasupa, P. Chatkamjuncharoen, C. Wuttilertdeshar, and M. Sugimoto, "Using image features and eye tracking device to predict human emotions towards abstract images," in *Proc. Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, 2015, pp. 419–430.

[48] W. Liu, W.-L. Zheng, and B.-L. Lu, "Emotion recognition using multimodal deep learning," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2016, pp. 521–529.

[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. Int'l Conf. Machine Learning (ICML)*, 2011, pp. 689–696.

[50] H. Tang, W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using deep neural networks," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2017, pp. 811–819.

[51] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[52] Z. Jia, Y. Lin, J. Wang, Z. Feng, X. Xie, and C. Chen, "HetEmotionNet: Two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition," in *Proc. ACM Int'l Conf. Multimedia (ACM MM)*, 2021, pp. 1047–1056.

[53] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2020, pp. 12546–12555.

[54] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artificial Intelligence*, 2020, vol. 34, pp. 12321–12328.

[55] D. Wang, G. Li, W. Jia, and X. Luo, "Saliency-driven scaling optimization for image retargeting," *The Visual Computer*, vol. 27, no. 9, pp. 853–860, 2011.

[56] D. Valdez-Balderas, O. Muraveynyk, and T. Smith, "Fast hybrid image retargeting," in *Proc. IEEE Int'l Conf. Image Processing (IEEE ICIP)*, 2021, pp. 1849–1853.

[57] S. Qian, Y. Shi, H. Wu, J. Liu, and W. Zhang, "An adaptive enhancement algorithm based on visual saliency for low illumination images," *Applied Intelligence*, vol. 52, no. 2, pp. 1770–1792, 2022.

[58] F. Fan, Y. Ma, J. Huang, and Z. Liu, "Infrared image enhancement based on saliency weight with adaptive threshold," in *Proc. IEEE Int'l Conf. Signal and Image Processing (IEEE ICSIP)*, 2018, pp. 225–230.

[59] Y. Dai, C. Xue, and L. Zhou, "Visual saliency guided perceptual adaptive quantization based on HEVC intra-coding for planetary images," *Plos One*, vol. 17, no. 2, pp. e0263729, 2022.

[60] Y. Patel, S. Appalaraju, and R. Manmatha, "Saliency driven perceptual image compression," in *Proc. IEEE/CVF Winter Conf. Applications of Computer Vision (IEEE/CVF WACV)*, 2021, pp. 227–236.

[61] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool, "The interestingness of images," in *Proc. IEEE/CVF Int'l Conf. Computer Vision (IEEE/CVF ICCV)*, 2013, pp. 1633–1640.

[62] Y. Li, P. Xu, D. Lagun, and V. Navalpakkam, "Towards measuring and inferring user interest from gaze," in *Proc. Int'l Conf. World Wide Web (WWW) Companion*, 2017, pp. 525–533.

[63] S. Bazrafkan, A. Kar, and C. Costache, "Eye gaze for consumer electronics: Controlling and commanding intelligent systems," *IEEE Consumer Electronics Magazine*, vol. 4, no. 4, pp. 65–71, 2015.

[64] Q. Zhao, S. Chang, M. Harper, and J. Konstan, "Gaze prediction for recommender systems," in *Proc. ACM Conf. Recommender Systems (ACM RecSys)*, 2016, pp. 131–138.

[65] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Estimation of user-specific visual attention based on gaze information of similar users," in *Proc. IEEE Global Conf. Consumer Electronics (IEEE GCCE)*, 2019, pp. 486–487.

[66] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[67] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE/CVF Int'l Conf. Computer Vision (IEEE/CVF ICCV)*, 2009, pp. 2106–2113.

[68] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. Int'l Conf. Computational Statistics (COMPSTAT)*, 2010, pp. 177–186.

[69] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark," http://saliency.mit.edu/.

[70] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2015, pp. 1072–1080.

[71] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "User-centric visual attention estimation based on relationship between image and eye gaze data," in *Proc. IEEE Global Conf. Consumer Electronics (IEEE GCCE)*, 2018, pp. 44–45.

[72] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "User-specific visual attention estimation based on visual similarity and spatial information in images," in *Proc. IEEE Int'l Conf. Consumer Electronics-Taiwan (IEEE ICCE-TW)*, 2019, pp. 479–480.

[73] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2017, pp. 4700–4708.

[74] T. V. Nguyen and E. V. Bonilla, "Collaborative multi-output Gaussian processes," in *Proc. Association for Uncertainty in Artificial Intelligence (UAI)*, 2014, pp. 643–652.

[75] H. Liu, J. Cai, and Y.-S. Ong, "Remarks on multi-output Gaussian process regression," *Knowledge-Based Systems*, vol. 144, pp. 102–121, 2018.

[76] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention," *Sensors*, vol. 20, no. 8:2170, pp. 1–15, 2020.

[77] S. Fan, Z. Shen, M. Jiang, B. Koenig, J. Xu, M. Kankanhalli, and Q. Zhao, "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2018, pp. 7521–7531.

[78] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOX: Exceeding YOLO series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[79] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artificial Intelligence*, 2017, vol. 31, pp. 4278–4284.

[80] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.

[81] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Few-shot personalized saliency prediction using person similarity based on collaborative multi-output Gaussian process regression," in *Proc. IEEE Int'l Conf. Image Processing (IEEE ICIP)*, 2021, pp. 1469–1473.

[82] R. Rothe, *A deep understanding from a single image*, Ph.D. thesis, ETH Zurich, 2016.

[83] Z. Xia, X. Wang, L. Zhang, Z. Qin, X. Sun, and K. Ren, "A privacy-preserving and copy-deterrence content-based image retrieval scheme in cloud computing," *IEEE Trans. Information Forensics and Security*, vol. 11, no. 11, pp. 2594–2608, 2016.

[84] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2016, pp. 779–788.

[85] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2015, pp. 3431–3440.

[86] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE/CVF Int'l Conf. Computer Vision (IEEE/CVF ICCV)*, 2015, pp. 1520–1528.

[87] D. Maturana and S. Scherer, "VoxNet: A 3D convolutional neural network for real-time object recognition," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IEEE IROS)*, 2015, pp. 922–928.

[88] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, vol. 22, no. 12, pp. 1349–1380, 2000.

[89] W. Wang and Q. He, "A survey on emotional semantic image retrieval," in *Proc. IEEE Int'l Conf. Image Processing (IEEE ICIP)*, 2008, pp. 117–120.

[90] P. Vuilleumier, "How brains beware: Neural mechanisms of emotional attention," *Trends in Cognitive Sciences*, vol. 9, no. 12, pp. 585–594, 2005.

[91] R. Compton, "The interface between emotion and attention: A review of evidence from psychology and neuroscience," *Behavioral and Cognitive Neuroscience Reviews*, vol. 2, no. 2, pp. 115–129, 2003.

[92] D. Tao, X. Li, X. Wu, and S. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, vol. 29, no. 10, pp. 1700–1715, 2007.

[93] K. Sugata, T. Ogawa, and M. Haseyama, "Emotion estimation via tensor-based supervised decision-level fusion from multiple Brodmann areas," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2017, pp. 999–1003.

[94] G. Huang, Q. Zhu, and C. Siew, "Extreme learning machine: A new learning scheme of feedforward neural networks," in *Proc. IEEE Int'l Joint Conf. Neural Networks (IEEE IJCNN)*, 2004, vol. 2, pp. 985–990.

[95] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified

framework for multi-label image classification," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2016, pp. 2285–2294.

[96] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Information Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[97] N. Karessli, Z. Akata, B. Schiele, and A. Bulling, "Gaze embeddings for zero-shot image classification," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2017, pp. 4525–4534.

[98] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3, pp. 321–377, 1936.

[99] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2016, pp. 2818–2826.

[100] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[101] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEE/CVFE Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2016, pp. 770–778.

[102] X. Tan, Y. Zhang, S. Tang, J. Shao, F. Wu, and Y. Zhuang, "Logistic tensor regression for classification," in *Proc. Int'l Conf. Intelligent Science and Intelligent Data Engineering*, 2012, pp. 573–581.

[103] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," *arXiv preprint arXiv: 1610.02357*, 2017.

[104] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[105] V. Vapnik, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, vol. 24, no. 6, pp. 774–780, 1963.

[106] B. Schuller, Z. Zhang, F. Weninger, and G. Rigoll, "Using multiple databases for training in emotion recognition: To unite or to vote?," in *Proc. ISCA Annual Conference of the Int'l Speech Communication Association (ISCA INTERSPEECH)*, 2011, pp. 1553–1556.

[107] H. Gunes and M. Piccardi, "Affect recognition from face and body: Early fusion vs. late fusion," in *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics (IEEE SMC)*, 2005, vol. 4, pp. 3437–3443.

[108] D. Ayata, Y. Yaslan, and M. E. Kamasak, "Emotion based music recommendation system using wearable physiological sensors," *IEEE Trans. Consumer Electronics*, vol. 64, no. 2, pp. 196–203, 2018.

[109] R. Sawata, T. Ogawa, and M. Haseyama, "Human-centered favorite music classification using EEG-based individual music preference via deep time-series CCA," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2021, pp. 1320–1324.

[110] S. N. Mohammed and A. K. A. Hassan, "A survey on emotion recognition for human robot interaction," *Journal of Computing and Information Technology*, vol. 28, no. 2, pp. 125–146, 2020.

[111] S. Nayak, B. Nagesh, A. Routray, and M. Sarma, "A human–computer interaction framework for emotion recognition through time-series thermal video sequences," *Computers & Electrical Engineering*, vol. 93, pp. 107280, 2021.

[112] R. W. Picard, *Affective computing*, MIT press, 2000.

[113] N. Liu, Y. Fang, L. Li, L. Hou, F. Yang, and Y. Guo, "Multiple feature fusion for automatic emotion recognition using EEG signals," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2018, pp. 896–900.

[114] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Human-centric emotion estimation based on correlation maximization considering changes with time in visual attention and brain activity," *IEEE Access*, vol. 8, pp. 203358–203368, 2020.

[115] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[116] H.-F. Jiang, X.-Y. Guan, W.-Y. Zhao, L.-M. Zhao, and B.-L. Lu, "Generating multimodal features for emotion classification from eye movement signals," in *Proc. Int'l Conf. Neural Information Processing (ICONIP)*, 2019, pp. 59–66.

[117] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Estimation of emotion labels via tensor-based spatiotemporal visual attention analysis," in *Proc. IEEE Int'l Conf. Image Processing (IEEE ICIP)*, 2019, pp. 4105–4109.

[118] M. Liu, Y. Fu, and T. S. Huang, "An audio-visual fusion framework with joint dimensionality reducton," in *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2008, pp. 4437–4440.

[119] Q.-S. Sun, Z.-d. Liu, P.-A. Heng, and D.-S. Xia, "A theorem on the generalized canonical projective vectors," *Pattern Recognition*, vol. 38, no. 3, pp. 449–452, 2005.

[120] K.-H. Pong and K.-M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556–567, 2014.

[121] E. H. El-Shazly, M. M. Abdelwahab, A. Shimada, and R.-i. Taniguchi, "Real time algorithm for efficient HCI employing features obtained from MYO sensor," in *Proc. Int'l Midwest Symposium on Circuits and Systems (MWSCAS)*, 2016, pp. 1–4.

[122] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE/CVF Int'l Conf. Computer Vision and Pattern Recognition (IEEE/CVF CVPR)*, 2016, pp. 478–487.

[123] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowledge and Data Engineering (IEEE TKDE)*, vol. 22, no. 10, pp. 1345–1359, 2010.

[124] A. Girouard, E. T. Solovey, L. M. Hirshfield, E. M. Peck, K. Chauncey, A. Sassaroli, S. Fantini, and R. J. Jacob, *From brain signals to adaptive interfaces: Using fNIRS in HCI*, pp. 221–237, 2010.

[125] Y. Hoshi, "Near-infrared spectroscopy for studying higher cognition," *Neural Correlates of Thinking*, pp. 83–93, 2009.

[126] D. Heger, R. Mutter, C. Herff, F. Putze, and T. Schultz, "Continuous recognition of affective states by functional near infrared spectroscopy signals," in *Proc. Humaine Association Conf. Affective Computing and Intelligent Interaction*, 2013, pp. 832–837.

[127] D. Bandara, S. Velipasalar, S. Bratt, and L. Hirshfield, "Building predictive models of emotion with functional near-infrared spectroscopy," *Int'l Journal of Human-Computer Studies*, vol. 110, pp. 75–85, 2018.

[128] T. Gruber, C. Debracque, L. Ceravolo, K. Igloi, B. Marin Bosch, S. Frühholz, and D. Grandjean, "Human discrimination and categorization of emotions in voices: A functional near-infrared spectroscopy (fNIRS) study," *Frontiers in Neuroscience*, vol. 14, pp. 570, 2020.

[129] Y. Kita, A. Gunji, K. Sakihara, M. Inagaki, M. Kaga, E. Nakagawa, and T. Hosokawa, "Scanning strategies do not modulate face identification: Eye-tracking and near-infrared spectroscopy study," *Plos One*, vol. 5, no. 6, pp. 1–10, 2010.

[130] Y. Suzuki, K. Shirahada, M. Kosaka, and A. Maki, "A new marketing methodology by integrating brain measurement, eye tracking, and questionnaire analysis," in *Proc. IEEE Int'l Conf. Service Systems and Service Management (IEEE ICSSSM)*, 2012, pp. 770–773.

[131] K. Fujiwara, N. Kiyota, K. Kunita, M. Yasukawa, K. Maeda, and X. Deng, "Eye movement performance and prefrontal hemodynamics during saccadic eye movements in the elderly," *Journal of Physiological Anthropology*, vol. 29, no. 2, pp. 71–78, 2010.

[132] M. Shensa, "The discrete wavelet transform: Wedding the a trous and mallat algorithms," *IEEE Trans. Signal Processing (IEEE TSP)*, vol. 40, no. 10, pp. 2464–2482, 1992.

[133] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[134] S. E. Leurgans, R. A. Moyeed, and B. W. Silverman, "Canonical correlation analysis when the data are curves," *Journal of Royal Statistical Society: Series B*, vol. 55, no. 3, pp. 725–740, 1993.

[135] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[136] G. Cai and B. Xia, "Convolutional neural networks for multimedia sentiment analysis," in *Proc. CCF Int'l Conf. Natural Language Processing and Chinese Computing (CCF NLPCC)*, 2015, pp. 159–167.

[137] M. Soleymani, M. Riegler, and P. Halvorsen, "Multimodal analysis of image search intent: Intent recognition in image search from user behavior and visual content," in *Proc. ACM Int'l Conf. Multimedia Retrieval (ICMR)*, 2017, pp. 251–259.

[138] S. Amemiya, H. Takao, and O. Abe, "Origin of the time lag phenomenon and the global signal in resting-state fMRI," *Frontiers in Neuroscience*, vol. 14, pp. 1–13, 2020.

[139] K. Hirasawa, K. Maeda, T. Ogawa, and M. Haseyama, "MvGAN maximizing time-lag aware canonical correlation for baseball highlight generation," in *Proc. IEEE Int'l Conf. Multimedia & Expo Workshops (IEEE ICMEW)*, 2020, pp. 1–6.

[140] Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Human emotion estimation using multi-modal variational autoencoder with time changes," in *Proc. IEEE Global Conf. Life Sciences and Technologies (IEEE LifeTech)*, 2021, pp. 67–68.

[141] M. Suzuki, K. Nakayama, and Y. Matsuo, "Joint multimodal learning with deep generative models," *arXiv preprint arXiv:1611.01891*, 2016.

[142] P. Sarma and S. Barma, "Review on stimuli presentation for affect analysis based on EEG," *IEEE Access*, vol. 8, pp. 51991–52009, 2020.

[143] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani, "Modular encoding and decoding models derived from Bayesian canonical correlation analysis," *Neural Computation*, vol. 25, no. 4, pp. 979–1005, 2013.

[144] C. Du, C. Du, L. Huang, and H. He, "Reconstructing perceived images from human brain activities with Bayesian deep multiview learning," *IEEE Trans. Neural Networks and Learning Systems (IEEE TNNLS)*, vol. 30, no. 8, pp. 2310–2323, 2018.

[145] C. Du, C. Du, and H. He, "Sharing deep generative representation for perceived image reconstruction from human brain activity," in *Proc. IEEE Int'l Joint Conf. Neural Networks (IEEE IJCNN)*, 2017, pp. 1049–1056.

[146] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "MVAE: Multimodal variational autoencoder for fake news detection," in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 2915–2921.

[147] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 2980–2988, 2015.

[148] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[149] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[150] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[151] T. A. Chang and B. K. Bergen, "Language model behavior: A comprehensive survey," *arXiv preprint arXiv:2303.11504*, 2023.

[152] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Computing Surveys (ACM CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[153] C. Zhang, Y. Xie, H. Bai, B. Yu, W. Li, and Y. Gao, "A survey on federated learning," *Knowledge-Based Systems*, vol. 216, pp. 106775, 2021.

[154] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint arXiv:2003.02133*, 2020.

List of Achievements

(A) **Journal**

[**A-1**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Tensor-based emotional category classification via visual attention-based heterogeneous CNN feature fusion," *Sensors* (2022IF: 3.9), vol. 20, no. 7:2146, 2020.

[**A-2**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Few-shot personalized saliency prediction based on adaptive image selection considering object and visual attention," *Sensors* (2022IF: 3.9), vol. 20, no. 8: 2170, 2020.

[**A-3**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Human-centric emotion estimation based on correlation maximization considering changes with time in visual attention and brain activity," *IEEE Access* (2022IF: 3.9), vol. 8, pp. 203358–203368, 2020.

[**A-4**] 諸戸祐哉, 前田圭介, 藤後廉, 小川貴弘, 長谷山美紀, "テキストおよび画像情報に基づく Focal Loss を導入した深層学習による冬期路面状態の分類," 土木学会 AI・データサイエンス論文集, vol. 3, no. J2, pp. 293–306, 2022.

[**A-5**] Ryota Goka, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa and Miki Haseyama, "Prediction of Shooting Events in Soccer Videos Using Complete Bipartite Graphs and Players' Spatial-temporal Relations," *Sensors* (2022IF: 3.9), vol. 23, no. 9: 4506, 2023.

[**A-6**] 諸戸祐哉, 前田圭介, 藤後廉, 小川貴弘, 長谷山美紀, "時系列データを用いた Multi-modal Transformer に基づく冬期路面状態の分類," 土木学会 AI・データサイエンス論文集, vol. 4, no. 3, pp. 402–413, 2023.

[**A-7**] Yuya Moroto, Yingrui Ye, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Zero-shot visual sentiment prediction via cross-domain knowledge distillation," *IEEE Open*

Journal of Signal Processing (IEEE OJSP, 2022IF: 2.8), 2024. (Accepted for publication)

(B) International Conference

[**B-1**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "User-centric visual attention estimation based on relationship between image and eye gaze data," in *Proc. the IEEE Global Conference on Consumer Electronics (IEEE GCCE)*, pp. 44–45, 2018.

[**B-2**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Estimation of visual attention via canonical correlation between visual and gaze-based features," in *Proc. the IEEE Global Conference on Life Sciences and Technologies (IEEE LifeTech)*, pp. 229–230, 2019.

[**B-3**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "User-specific visual attention estimation based on visual similarity and spatial information in images," in *Proc. the IEEE International Conference on Consumer Electronics - Taiwan (IEEE ICCE-TW)*, pp. 479–480, 2019.

[**B-4**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Estimation of emotion labels via tensor-based spatiotemporal visual attention analysis," in *Proc. the IEEE International Conference on Image Processing (IEEE ICIP)*, pp. 4105–4109, 2019.

[**B-5**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Estimation of user-specific visual attention based on gaze information of similar users," in *Proc. the IEEE Global Conference on Consumer Electronics (IEEE GCCE)*, pp. 486–487, 2019.

[**B-6**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Estimation of person-specific visual attention via selection of similar persons," in *Proc. the IEEE International Conference on Consumer Electronics - Taiwan (IEEE ICCE-TW)*, pp. 1–2, 2020.

[**B-7**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Estimation of user-specific visual attention considering individual tendency toward gazed objects," in *Proc. the IEEE Global Conference on Consumer Electronics (IEEE GCCE)*, pp. 554–555, 2020.

[**B-8**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Human emotion estimation using multi-modal variational autoencoder with time changes," in *Proc. the IEEE Global Conference on Life Sciences and Technologies (IEEE LifeTech)*, pp. 82–83, 2021.

[**B-9**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Few-shot personalized saliency prediction using person similarity based on collaborative multi-output gaussian process regression," in *Proc. the IEEE International Conference on Image Processing (IEEE ICIP)*, pp. 1469–1473, 2021.

[**B-10**] Yingrui Ye, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Visual sentiment prediction using few-shot learning via distribution relations of visual features," in *Proc. the IEEE Global Conference on Consumer Electronics (IEEE GCCE)*, pp. 217–218, 2021.

[**B-11**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Human emotion recognition using multi-modal biological signals based on time lag-considered correlation maximization," in *Proc. the IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, pp. 4683–4687, 2022.

[**B-12**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Few-shot personalized saliency prediction with similarity of gaze tendency using object-based structural information," in *Proc. the IEEE International Conference on Image Processing (IEEE ICIP)*, pp. 3823–3827, 2022.

[**B-13**] Yingrui Ye, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Visual sentiment prediction using cross-way few-shot learning based on knowledge distillation," in *Proc. the IEEE International Conference on Image Processing (IEEE ICIP)*, pp. 3838–3842, 2022.

[**B-14**] Ryota Goka, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Shoot event prediction from soccer videos by considering players' spatio-temporal relations," in *Proc. the IEEE Global Conference on Consumer Electronics (IEEE GCCE)*, pp. 417–418, 2022.

[**B-15**] Yingrui Ye, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Affective embedding framework with semantic representations from tweets for zero-shot visual sentiment prediction," in *Proc. the ACM Multimedia Asia (ACM MM Asia)*, pp. 1–7, 2022.

[**B-16**] Ryota Goka, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Prediction of shoot events by considering spatio-temporal relations of multimodal features," in *Proc. the IEEE International Conference on Consumer Electronics - Taiwan (IEEE ICCE-TW)*, pp. 793–794, 2023.

[**B-17**] Ryota Goka, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Shoot Event Prediction in Soccer Considering Expected Goals Based on Players' Positions," in *Proc. the IEEE International Conference on Consumer Electronics - Taiwan (IEEE ICCE-TW)*, pp. 449–450, 2023.

[**B-18**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama, "Multi-view variational recurrent neural network for human emotion recognition using multi-modal biological signals," in *Proc. the IEEE International Conference on Image Processing (IEEE ICIP)*, pp. 2925–2929, 2023.

[**B-19**] Yuya Moroto*, Rintaro Yanagi*, Naoki Ogawa, Kyohei Kamikawa, Keigo Sakurai, Ren Togo, Keisuke Maeda, Takahiro Ogawa, Miki Haseyama, "Personalized content recommender system via non-verbal interaction using face mesh and facial expression," in *Proc. the ACM Multimedia (ACM MM), Demos and Videos Track*, pp. 9399–9401, 2023. *: Equal Contributions

[**B-20**] Ryota Goka, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, Huang-Chia Shih and Miki Haseyama, "Masked modeling-based action event prediction considering bidirectional time-series in soccer," in *Proc. the International Workshop on Advanced Image Technology (IWAIT)*, 2023. (Accepted for publication)

[**B-21**] Yuya Moroto, Yingrui Ye, Keisuke Maeda, Takahiro Ogawa, Huang-Chia Shih and Miki Haseyama, "Zero-shot visual sentiment prediction via cross-domain knowledge distilla-

tion," in *Proc. the IEEE International Conference on Acoustics, Speech and Signal Processing (IEEE ICASSP)*, 2024. (IEEE OJSP，2024 の内容として発表予定)

(C) Domestic Conference (Technical Report & Lecture)

[**C-1**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "画像注視時の注視領域の時間変化を考慮したテンソル解析に基づく感情推定に関する検討," 平成 30 年度電気・情報関係学会北海道支部連合大会, pp. 137–138, 2018.

[**C-2**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "画像の視覚的および空間的特徴に基づくユーザに特化した注視領域推定の高精度化に関する検討 ～視覚的特徴の類似度と推定精度の関係性に関する一考察～," イメージ・メディア・クオリティ研究会 (IMQ), pp. 13–16, 2019.

[**C-3**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "視線情報を考慮した画像のテンソル表現に基づく感情ラベル推定に関する検討 –複数ユーザの推定結果の統合に基づく高精度化–," 第 22 回 画像の認識・理解シンポジウム (MIRU), pp. 1–4, 2019.

[**C-4**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "Sparse Bayesian Learning に基づく注視領域の時間変化を考慮したヒトの感情推定に関する検討," 令和元年度電気・情報関係学会北海道支部連合大会, pp.149–150, 2019.

[**C-5**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "画像注視時のヒトの感情推定のための視線特徴の推定に関する検討," 映像情報メディア学会技術報告, vol. 44,no. 6, pp. 85–89, 2020.

[**C-6**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "路面画像を用いた異常検知に基づく路面状態の識別に関する検討," 令和 2 年度電気・情報関係学会北海道支部連合大会, pp. 118–119, 2020.

[**C-7**] 叶 穎睿, 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "Few-shot learning を用いた感情ラベル推定における複数のデータセット利用に関する初期検討," 令和 3 年度電気・情報関係学会北海道支部連合大会, pp.123–124, 2021.

[**C-8**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, “［特別講演］路面画像を用いた深層学習に基づく路面状態の識別に関する検討,” 映像情報メディア学会技術報告, vol. 45, no. 4, pp. 165–169, 2021.

[**C-9**] 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, “画像中の物体情報を考慮したユーザ類似度に基づく個人に特化した注視領域の推定に関する検討,” 映像情報メディア学会技術報告, vol. 44, no. 6, pp. 181–186, 2022.

[**C-10**] 叶 穎睿, 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, “知識蒸留を用いた Few-shot Learning に基づく画像の感情ラベル推定に関する検討,” 映像情報メディア学会技術報告, vol. 46, no. 6, pp. 171–175, 2022.

[**C-11**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawaand Miki Haseyama, “Few-shot personalized saliency prediction via person similarity usingtensor-based regression,” 第 25 回 画像の認識・理解シンポジウム (MIRU), pp. 1–4, 2022.

[**C-12**] 五箇 亮太, 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, “サッカー競技のスカウティング映像における選手間の時空間的関係を考慮した不確実性に基づくシュートイベント予測に関する検討,” 令和 4 年度電気・情報関係学会北海道支部連合大会, pp. 196–197, 2022.

[**C-13**] 諸戸 祐哉, 前田 圭介, 藤後 廉, 小川 貴弘, 長谷山 美紀, “テキストおよび画像情報に基づく Focal Loss を導入した深層学習による冬期路面状態の分類,” 第 3 回 AI・データサイエンスシンポジウム, 2022. (土木学会 AI・データサイエンス論文集, 2022 の内容として発表)

[**C-14**] 五箇 亮太, 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, “サッカー映像における時空間的関係を考慮したシュート予測の高精度化に関する検討 ～競技者のチーム情報に基づく完全二部グラフの導入～,” 映像情報メディア学会技術報告, vol. 47, no. 6, pp. 227–232, 2023.

[**C-15**] Yingrui Ye, Yuya Moroto, Keisuke Maeda, Takahiro Ogawaand Miki Haseyama, “Zero-shot visual sentiment prediction with cross-domain sentiments using knowledge distillation,” 第 26 回画像の認識・理解シンポジウム (MIRU 2023), pp. 1–5, 2023.

[**C-16**] Yuya Moroto, Keisuke Maeda, Takahiro Ogawaand Miki Haseyama, "Human emotion recognitionwhile viewing images based on multi-viewvariational recurrent neural network," 第 26 回画像の認識・理解シンポジウム (MIRU 2023), pp. 1–5, 2023.

[**C-17**] 諸戸 祐哉, 前田 圭介, 藤後 廉, 小川 貴弘, 長谷山 美紀, "時系列データを用いた Multi-modal Transformer に基づく冬期路面状態の分類," 第 4 回 AI・データサイエンスシンポジウム, 2023. (土木学会 AI・データサイエンス論文集, 2023 の内容として発表)

[**C-18**] 五箇 亮太, 諸戸 祐哉, 前田 圭介, 小川 貴弘, 長谷山 美紀, "双方向 Transformer に基づいたサッカー選手のイベントデータからの行動推定に関する検討," 映像情報メディア学会技術報告, 2024. (発表予定)

(D) **Award**

[**D-1**] IEEE GCCE 2018 Outstanding Paper Award, 2018.

[**D-2**] 2nd Prize IEEE LifeTech 2019 Excellent Paper Award, 2019.

[**D-3**] The 2019 IEEE Sapporo Section Student Paper Contest Encouraging Prize, 2020.

[**D-4**] 令和 2 年度 電気・情報関係学会北海道支部連合大会 若手優秀論文発表賞, 2020.

[**D-5**] IEEE LifeTech 2021 Excellent Poster Award for On-site Poster Presentation, 2021.

[**D-6**] 令和 2 年度 電子情報通信学会北海道支部 学生奨励賞, 2021.

[**D-7**] The 2021 IEEE Sapporo Section Encouragement Award, 2022.

[**D-8**] The 2021 IEEE Sapporo Section Student Paper Contest Best Presentation Award, 2022.

[**D-9**] 令和 4 年度 電気・情報関係学会北海道支部連合大会 若手優秀論文発表賞, 2022.

[**D-10**] IEEE ICCE-TW 2023 Best Paper Award Honorable Mention, 2023.

[**D-11**] AI・データサイエンス奨励賞 Intelligence, Informatics and Infrastructure Award for Outstanding Potential Paper, 2023.

[**D-12**] IWAIT2024 Best Paper Award, 2024.

(E) **Scholarship**

[**E-1**] 日本学生支援機構 第一種奨学金 特に優れた業績による奨学金返還免除（全額免除）, 2019.4-2021.3.

[**E-2**] 似鳥国際奨学財団 北海道みらいIT人財奨学金, 2020.4-2022.3.

[**E-3**] 日本学術振興会 特別研究員 DC1, 2021.4-2024.3.

[**E-4**] JEES・三菱商事科学技術学生奨学金, 2022.4-2023.3.

[**E-5**] Sky 大浦 ICT 奨学財団, 2023.4-2024.3.