Do Smart Glasses Dream of Sentimental Visions? Deep *Emotionship* Analysis for Eyewear Devices

YINGYING ZHAO[*] and YUHU CHANG[*], School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
YUTIAN LU, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
YUJIANG WANG[†], Department of Computing, Imperial College London, United Kingdom
MINGZHI DONG, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
QIN LV, Department of Computer Science, University of Colorado Boulder, United States
ROBERT P. DICK, Department of Electrical Engineering and Computer Science, University of Michigan, United States
FAN YANG, School of Microelectronics, Fudan University, China
TUN LU, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
NING GU, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
LI SHANG, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China

[*] Equal contribution
[†] Corresponding author

Authors' addresses: Yingying Zhao, yingyingzhao@fudan.edu.cn; Yuhu Chang, yhchang14@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Yutian Lu, 20210240098@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Yujiang Wang, yujiang.wang14@imperial.ac.uk, Department of Computing, Imperial College London, London, United Kingdom; Mingzhi Dong, mingzhidong@gmail.com, School of Computer Science, Fudan University, Shanghai, China, 200438, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Qin Lv, qin.lv@colorado.edu, Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, United States, 80309; Robert P. Dick, dickrp@umich.edu, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, United States, 48109; Fan Yang, yangfan@fudan.edu.cn, School of Microelectronics, Fudan University, Shanghai, China, 201203; Tun Lu, lutun@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Ning Gu, ninggu@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Li Shang, lishang@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438, Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438.

2474-9567/2022/3-ART38 $15.00
https://doi.org/10.1145/3517250

Emotion recognition in smart eyewear devices is valuable but challenging. One key limitation of previous works is that the expression-related information like facial or eye images is considered as the only evidence of emotion. However, emotional status is not isolated; it is tightly associated with people's visual perceptions, especially those with emotional implications. However, little work has examined such associations to better illustrate the causes of emotions. In this paper, we study the *emotionship* analysis problem in eyewear systems, an ambitious task that requires classifying the user's emotions and semantically understanding their potential causes. To this end, we describe *EMOShip*, a deep-learning-based eyewear system that can automatically detect the wearer's emotional status and simultaneously analyze its associations with semantic-level visual perception. Experimental studies with 20 participants demonstrate that, thanks to its awareness of *emotionship*, *EMOShip* achieves superior emotion recognition accuracy compared to existing methods (80.2% vs. 69.4%) and provides a valuable understanding of the causes of emotions. Further pilot studies with 20 additional participants further motivate the potential use of *EMOShip* to empower emotion-aware applications, such as emotionship self-reflection and emotionship life-logging.

CCS Concepts: • **Human-centered computing → Mobile devices**.

Additional Key Words and Phrases: Smart Eyewear System, Emotionship, Emotion Recognition, Sentiment Analysis, Image Captioning, Visual Question Answering

ACM Reference Format:
Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. 2022. Do Smart Glasses Dream of Sentimental Visions? Deep *Emotionship* Analysis for Eyewear Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 38 (March 2022), 29 pages. https://doi.org/10.1145/3517250

1 INTRODUCTION

Research in social science and psychology indicates that our emotional state can considerably affect several aspects of daily life, including thoughts and behaviors [13], decision making [49], cognitive focuses [18], performance on assessments [46], physical health [17], and mental well-beings [57]. Given the significant impacts of emotions, emotion recognition is one of the most crucial research topics in affective computing [45], and it can be applied to a wide range of human-computer interaction (HCI) scenarios to improve user experience. Intelligent eyewear systems are especially well suited to carry out and benefit from emotion recognition.

A common goal of smart eyewear devices is to deliver intelligent services with personalized experiences. This requires understanding the users, especially their affective status. As indicated by previous studies [4, 13–15, 67], the ability to recognize emotion can greatly enhance user experience in various HCI scenarios. More importantly, an emotion-sensitive wearable front-end would enable a variety of personalized back-end applications, such as emotional self-reflection [18, 23], emotional life-logging [6], emotional retrieving and classification [64], and mood tracking [56].

Recognizing emotions using smart eyewear devices is challenging. The majority of state-of-the-art emotion recognition techniques [19, 30, 32, 34, 35] use deep learning models to classify expressions from full facial images. However, it is typically difficult to capture the entire face using sensors that can economically be integrated into current eyewear devices. This mismatch between economical sensors and analysis techniques hinders the practical application of existing emotion recognition methods in eyewear.

To address this challenging problem, previous works [21, 41, 51] adopted engineering-based approaches to extract hand-crafted features from eye regions instead of the whole facial images to compute the affective status. With the embedding of eye-tracking cameras in commercial eyewear devices, recent eyewear systems developed convolutional neural networks (CNN) to extract deep affective features from eye-camera-captured images (typically eye regions) for head-mounted virtual reality (VR) glasses [27] and smart glasses [61]. These prior works have limited recognition accuracy; they predict human emotions based entirely on the portions of facial expressions visible near the eyes, ignoring the subtle yet crucial associations between emotional status and visual perception. In fact, information about the emotional state can be found in both expressions and visual experiences. The additional sentimental clues provided by visual experience improve emotion recognition compared to considering facial expressions alone.

As shown in studies of behaviors and neuroscience [12, 43, 44], sentimental content in scenes is generally prioritized by people's visual attention compared to emotionally neutral content. Emotion-arousing content is known as *emotional stimuli*. For example, viewing a child playing with parents can lead to joyfulness, while a crying woman who just lost her husband can lead to sadness. In other words, emotion is not an isolated property; instead, it is tightly connected with visual emotional stimuli. Emotions can be closely associated with varying sentimental visual perception, especially for eyewear devices with rapidly altering scenes.

Based on the above observations, we study the *emotionship* analysis problem in eyewear devices. The term *emotionship* denotes the association of emotional status with the relevant hints in expression and visual attention. Through *emotionship* analysis, we aim to recognize emotions with better accuracy and understand the semantic causes [1]. Quantitative measurement of visual attention will be used to better estimate emotional state. In this paper, we adopt the widely accepted emotion categorization system that classifies emotions into six basic categories [20] plus neutrality (following [61]) to define the status of emotions. It is important to note that a semantic-level understanding of visual experiences is necessary, since a certain attention region may consist of multiple objects, leaving the associations among emotions and objects ambiguous. In other words, we need to capture the semantic attributes of visual perceptions. Compared with traditional emotion recognition techniques, the proposed *emotionship* analysis is arguably more ambitious and more difficult, as additional challenges arise from the semantic analysis of the human visual perception, its association with the emotional status, etc. However, a successful *emotionship* analysis framework will clearly lead to a truly personalized eyewear system that is capable of performing unseen and valuable emotion-aware downstream tasks.

In this work, we present such an *emotionship*-aware eyewear system for the first time. As shown in Fig. 1, our eyewear system, called *EMOShip*, is a deep learning system capable of recognizing the semantic attributes of the visual attentive regions, the expression-related information in eye images, and the emotional states based on both sources of information. At the heart of *EMOShip* is *EMOShip*-Net, a deep neural network designed to address the new challenges in *emotionship* analysis. To extract the semantic attributes of visual perceptions, we combine gaze points from eye-tracking [29] with the visual features model VinVL [66] plus a vision-language (VL) fusion model OSCAR+ [66]. The sentimental clues in visual perceptions are synthesized with the expression-related information in eye images to predict emotional status more accurately and robustly. The contributions of visual perception to emotional states, which are subtle and challenging to measure, are quantitatively evaluated by a Squeeze-and-Excitation (SE) network [28] that fuses the scene and eye features. To evaluate the in-lab performance of *EMOShip*, we collect and construct a new dataset named EMO-Film. With the availability of visual perceptions' semantic attributes, the emotional states, and the emotional impacts of visual attentions, our smart glasses system *EMOShip* outperforms baseline methods on the EMO-Film dataset in terms of emotion recognition accuracy, and more importantly, *EMOShip* provides a semantic understanding of the potential cause of such emotions. In-field pilot studies have been conducted to illustrate the capabilities of this *emotionship*-aware eyewear system and demonstrate its potential applications to a number of *emotionship*-relevant tasks such as emotionship self-reflection and emotionship life-logging.

In summary, this paper makes the following contributions.

(1) This work describes the design of smart eyewear, called *EMOShip*, that measures the relationship between the semantic attributes of visual attention and the emotional state of the wearer. Using this learned relationship increases the accuracy of emotional state recognition.

[1] In this paper, the phrase "understanding the cause" refers to the understanding of the momentary associations between emotional states and the scene images in eyewear devices. It should be distinguished with the understanding of the emotions' causality [11] which is a different task.

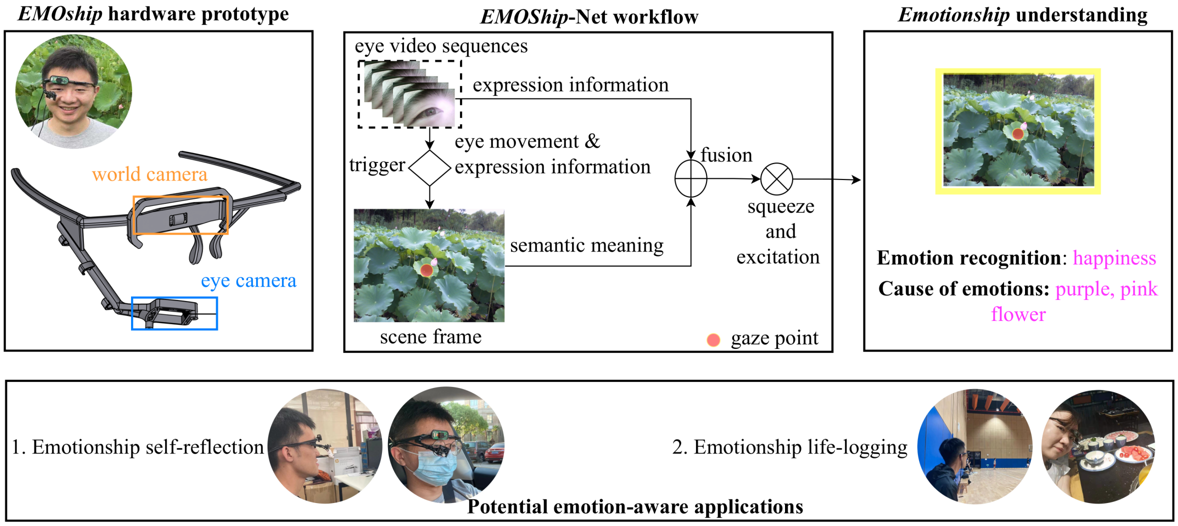

EMOship hardware prototype
world camera
eye camera
EMOShip-Net workflow
eye video sequences
expression information
eye movement &
expression information
trigger
fusion
squeeze
and
excitation
semantic meaning
scene frame
gaze point
Emotionship understanding
Emotion recognition: happiness
Cause of emotions: purple, pink flower
1. Emotionship self-reflection
2. Emotionship life-logging
Potential emotion-aware applications

Fig. 1. The proposed *EMOShip* smart eyewear system.

(2) The *EMOShip* eyewear is equipped with a deep neural network *EMOShip*-Net that extracts expression-related affective features and sentimental clues in visual attention, fusing them to more accurately recognition emotion and quantify *emotionship* associations.
(3) On the self-collected EMO-Film dataset, *EMOShip* achieves approximately 10.8% higher emotion recognition accuracy than the baseline methods, and demonstrates the potential to provide valuable sentimental clues for *emotionship* understanding.
(4) We perform in-field pilot studies on two inspiring down-stream applications – emotionship self-reflection and emotionship life-logging, to illustrate potential uses of *EMOShip*. This three-week study of 20 participants, shows that *EMOShip* captures emotional moments with a precision of 82.8%. A survey-based study shows that 16 out of 20 users found emotionship self-reflection beneficial, while 15 out of 20 users found emotion life-logging beneficial.

The rest of the paper is organized as follows. Section 2 surveys related work. Section 3 describes the algorithms and system implementation of *EMOShip*. Section 4 presents the experimental results. Section 5 describes in-field pilot studies. Section 6 concludes.

2 RELATED WORK

This section surveys related work in the field of (1) emotion recognition, (2) image sentiment analysis, (3) vision-language models, and (4) gating mechanisms. We also highlight the key contributions of *EMOShip* compared to prior work.

2.1 Emotion Recognition

Ekman proposed a well-known and widely adopted emotion categorization system that divided emotions into six basic categories: happiness, sadness, fear, anger, disgust, and surprise [20]. *Neutrality* denotes an additional, seventh, state: the absence of emotion [61]. The seven basic emotions are widely accepted [1, 61]: we adopt this emotion categorization system.

Most recent works involve deep models to classify the seven basic emotions from whole-face images [19, 30, 32, 34, 35], as facial expressions are one of the most common channels for humans to express emotions [22]. It is difficult for smart eyewear devices to capture whole-face images but eye-region images have been shown to contain sufficient expression-related information [61] for emotion estimation, and these images can be easily captured using commodity eye cameras. Therefore, eye image analysis techniques are promising for emotion recognition in eyewear systems. Tarnowski et al. proposed to use eye-tracking information, mainly regarding eye movements and pupil diameters, for emotion recognition [55]. Aracena et al. presented an emotion recognition method based only on pupil size and gaze position [3]. Later, Wu et al. proposed a deep-learning-based network that extracts emotional features from single-eye images and uses them to classify emotional state [61].

In this work, we take one further step and study *emotionship* analysis, which considers users' facial expressions and their visual perceptions.

2.2 Image Sentiment Analysis

Visual sentiment analysis aims to predict the emotional states produced by images. This work mainly investigates visual sentiment analysis based on categorical approaches that divide the intended emotions from images into six categories [53], which is usually consistent with the emotion categorization system [20].

Early sentiment prediction used hand-crafted features to recognize intended emotions. Those features included color variance, composition, and image semantics [40]. Recently, numerous deep convolutional neural network (CNN) learning based sentiment prediction approaches have been proposed to extract deep features for sentiment prediction [53, 63]. Campos et al. conducted extensive experiments and compared the performance of several fine-tuned CNNs for visual sentiment prediction [7]. Zhu et al. proposed a unified CNN-RNN model to predict image emotions based on both low-level and high-level features by considering the dependencies of the features [70]. Rao et al. classified image emotions based on a proposed multi-level deep neural network that combined the local emotional information from emotional regions with global information from the whole image [48]. Yang et al. proposed a weakly supervised coupled convolutional network to provide effective emotion recognition by utilizing local information in images [63]. Later, they extended the proposed weakly supervised detection framework through a more detailed analysis for visual sentiment prediction [53].

The above works provide idea of image-based emotion estimation; we extract the sentimental features in scene images through a Vision-Language (VL) model.

2.3 Vision-Language Models

Vision-Language (VL) models are a relatively new field in computer vision [9, 33, 37, 54, 69]. VL models usually consist of two stages: 1. An object detection model is used to predict the Regions of Interest (RoIs) of each object and also to extract the feature embedding for each RoI, 2. A cross-modal fusion model is used to generate short descriptions of each RoI's semantic attributes. Therefore, a successful VL model will generate all RoIs in a scene image, the feature embedding for each RoI, and also the semantic attributes of each RoI. VinVL model [66] improves the performance of the vision module to extract visual presentations at higher qualities, and employs OSCAR [36], which is based on a transformer [58], to perform the cross-modal semantic attributes predictions. It is shown [66] that using VinVL features and training on multiple datasets can significantly improve the performance of the original OSCAR on a variety of downstream Natural Language Processing (NLP) tasks, and therefore the learned Vision-language fusion model is named as OSCAR+. The VinVL model [66] has achieved state-of-the-art performance in VL tasks, and the performance of its proposed OSCAR+ has also surpassed that of others on downstream NLP tasks.

Inspired by recent progress in VL models and the requirements of semantic understanding in *emotionship* analysis, we have adopted VinVL [66] and its proposed OSCAR+ [66] in *EMOShip*. The benefits of using VinVL and OSCAR+ are threefold. First, the semantic attributes of RoIs can be predicted, which can enable *emotionship*-awareness. Another advantage is that the semantic features of RoIs are also provided in VinVL. These features encode sufficient sentimental clues for fusion with eye-expression-related information to achieve more accurate emotion prediction. Last but not least, we are able to perform language analysis tasks like Question Answering (QA) through using OSCAR+, which allows our eyewear system to capture the summary tag of a visual region. To the best of our knowledge, we are the first to integrate a VL and NLP model into an eyewear system, making it aware of semantic attributes.

2.4 Gating Mechanisms

Gating mechanisms [2] [5, 38, 58, 59] spend more resources on more informative parts of the input data. Typically, for an input signal, the importance of each of its datum is weighted through a gating model, producing appropriately weighted output signals. There are a variety of gating models like the transformer [58] and non-local network [59], which are widely utilized in different fields like lip-reading [39] and image captioning [62]. Among those gating models, Squeeze-and-Excitation (SE) network [28] is most closely related to our smart glasses *EMOShip*. For a deep feature, SE network can learn the pattern of importance degree in a channel-wise manner, generating a scaling vector that adjusts each feature channel. In *EMOShip*, SE is employed when fusing the semantic features from VL models and eye features to predict the emotional state, and more importantly, to learn the emotional impacts from scene images.

3 SYSTEM DESIGN

This section presents *EMOShip* system design. It first defines the *emotionship* analysis problem, highlights the corresponding challenges, and then presents *EMOShip*-Net, the proposed deep *emotionship* analysis network, and describes the design and operation of the *EMOShip* software-hardware system.

3.1 Problem Definition

Emotion recognition methods for eyewear devices aim to identify the emotional state from expressions, typically using eye images. There can be various criteria regarding the emotional state, and we adopt a widely accepted standard [20, 61]. Specifically, the emotion is discretely classified into six basic categories [20] – happiness, surprise, anger, fear, disgust, and sadness. In addition, we employ neutrality to represent the absence of emotions [61]. Let $e^t \in \{0, 1, 2, 3, 4, 5, 6\}$ represent the emotional state at time step t and let $\mathbf{E}^t \in \mathbb{R}^{H_1 \times W_1 \times 3}$ be the eye images with height H_1 and width W_1. Recent smart eyewear devices [27, 61] utilized a deep network $\mathcal{N}_{eye}$ to obtain emotional predictions from eye images, i.e., $e^t = \mathcal{N}_{eye}(\mathbf{E}^t)$.

In this work, we aim to solve the *emotionship* analysis problem for eyewear devices, a task that is related to expression-based emotion recognition but is more ambitious. In particular, emotional state is inferred using both eye images and visual perceptions, and the impacts of visual perceptions on this emotional state, i.e., *emotionship*, are quantitatively evaluated. Since the visual attentive region usually covers multiple semantic objects, the semantic attributes of the visual perceptions should be distinguished to avoid confusion of those objects.

Let $\mathbf{I^t} \in \mathbb{R}^{H_2 \times W_2 \times 3}$ represent the scene image with height H_2 and width W_2. The goal is to determine the user's visual attentive region, or Region of Interest (RoI). We denote this RoI as $\mathbf{r}^t \in \mathbb{R}^4$, and $\mathbf{r}^t$ can be described as a rectangular area $(x_r^t, y_r^t, w_r^t, h_r^t)$, where (x_r^t, y_r^t) is the central point of the rectangle, and (w_r^t, h_r^t) denote the width and height of the rectangle, respectively, i.e., $\mathbf{r}^t \in \mathbb{R}^4$. The visual perceptions, denoted as $\mathbf{I}_{att}^t$, are obtained by cropping the region $\mathbf{r}^t$ out of $\mathbf{I}^t$. In contrast to existing emotion recognition methods, we aim to determine

[2] Note that gating mechanisms are more commonly known as *attention mechanisms*. However, to avoid confusion with the visual attention concept, we use the term *gating mechanism* in this work.

the emotional state e^t from both the visual perceptions $\mathbf{I}^t_{att}$ and eye images $\mathbf{E}^t$ through a deep model $\mathcal{N}_1$, i.e., $e^t = \mathcal{N}_1(\mathbf{I}^t_{att}, \mathbf{E}^t)$.

In addition to e^t, we also want to determine the impacts of visual perceptions $\mathbf{I}^t_{att}$ on this emotional state, i.e., the degree to which this emotion be attributed to visual experience. We define this impact as an influence score $IS^t \in [0, 1)$ which can be computed by inferring from $\mathbf{I}^t_{att}$, $\mathbf{E}^t$, and e^t. Assuming a deep model $\mathcal{N}_2$ is utilized, this can be written as $IS^t = \mathcal{N}_2(\mathbf{I}^t_{att}, \mathbf{E}^t, e^t)$. Intuitively, a larger IS^t score indicates that e^t is more associated with what the user observes, and vice versa for a smaller value.

The awareness of emotional state e^t and the influence score IS^t is not sufficient to fully reveal the *emotionship*, as we still need to understand the semantic attributes of visual attentions to unambiguously describe the potential cause for e^t. The semantic attribute is defined as a summary tag of the attentive region $\mathbf{I}^t_{att}$, e.g., "white, warm beaches" if $\mathbf{I}^t_{att}$ depicts a white beach in summer. We denote this summary tag as $\mathbf{s}^t$ and it clarifies the semantic cause for e^t at an abstract level, which is typically overlooked in previous works.

Let $\mathbf{ES}^t$ represent the *emotionship* and it can be formulated as

$$\mathbf{ES}^t = (e^t, \mathbf{I}^t_{att}, \mathbf{s}^t, IS^t). \tag{1}$$

In contrast with traditional emotion recognition, which isolates a user's emotional state from their surroundings and predict only e^t, our *emotionship* $\mathbf{ES}^t$ additionally encodes the potential causes for e^t, i.e., visual perceptions $\mathbf{I}^t_{att}$ with semantic attributes $\mathbf{s}^t$, while the degrees of their emotional influences are also indicated by IS^t. Awareness of *emotionship* can enable eyewear devices to understand the semantic causes of emotions and also learn how visual experiences affect emotions in a personalized manner. However, there are a number of challenges.

3.2 Challenges

Emotionship analysis faces three primary challenges.

The first is how to appropriately discover the semantic attributes of visual attention. With the embedding of the forward-facing world camera, smart eyewear devices can already estimate the gaze points [8] using eye-tracking techniques. Gaze points can be used to track human attention. However, knowing merely the gaze point is insufficient, as there can be multiple semantic objects near this point that may influence the user's emotional state. To avoid ambiguity, we must clearly identify the semantic meanings in the vicinity of the gaze point. In other words, the semantic summary tag $\mathbf{s}^t$ of the visual perceptions $\mathbf{I}^t_{att}$ is necessary, yet $\mathbf{s}^t$ can be challenging to obtain, especially for eyewear devices. We take inspiration from recent progress in visual features models [66] to extract the tag $\mathbf{s}^t$, as described in Section 3.3.

After the visual attentive regions have been located with semantic understandings, another challenge remains. How can the associations between human visual attention and emotional state be determined? The reason for emotional changes can be subtle and difficult to determine. They may be associated with sentimental visual perceptions, e.g., when a user observes that a child is playing with parents and then cheers up, we can reasonably assume that the happiness is caused by this scene. However, sentimentally neutral visions can also proceed, but not cause sudden emotional changes. Therefore, it is crucial to correctly identify the emotional contribution of visual attention, i.e., to compute its influence score IS^t. In our workflow, IS^t is automatically and implicitly learned using deep models, as described in Section 3.3.

There is one more challenge facing the prediction of emotional state e^t. Sentimental information in visual perceptions indeed provides insights on the potential causes of emotions, but is not reliable enough (alone) to recognize emotion. Therefore, the utilization of expression-related information, e.g., eye-images [61], remains valuable. In other words, we infer the emotional state e^t from both the sentimental clues in visual perceptions and the expression-related information in eye images, leading to more robust emotion recognition performance. We have incorporated the Squeeze-and-Excitation network [28] to fuse them together, as described in Section 3.3.

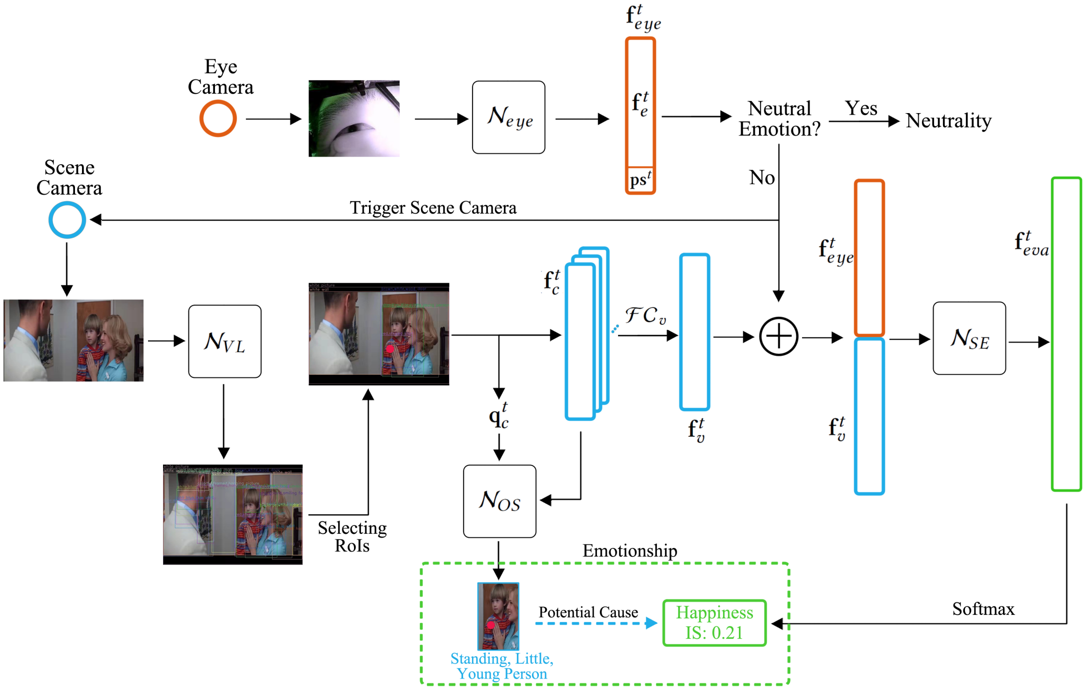


Fig. 2. The workflow of *EMOShip*-Net (Best Seen in Color).

To address the aforementioned challenges of *emotionship* analysis, we have devised a deep network named *EMOShip*-Net, the workflow of which is described in Section 3.3.

3.3 *EMOShip*-Net Workflow

3.3.1 Overall Pipeline. Fig. 2 illustrates the workflow of *EMOShip*-Net. At time step t, the input of the network contains two video streams: the eye images $\mathbf{E}^t$ taken by an inward-facing eye camera, and the scene images $\mathbf{I^t}$ recorded by another forward-facing world camera. The eye camera keeps tracking eye images $\mathbf{E}^t$ and monitors the rough emotional state, i.e., neutral or non-neutral. When a non-neutral emotion is spotted, the scene camera will be triggered to record scene $\mathbf{I^t}$. A vision-language (VL) model [66] is applied to extract all potential regions of interest (RoIs) with semantic tags in $\mathbf{I^t}$. The visual attentive region is determined from those RoIs based on the gaze point, and the summary tag for the selected area is obtained by a Question Answering (QA) process using the OSCAR+ [36] vision-language fusion model. The features of the attentive regions, which are also provided by the VL model, are fused with the eye features using a Squeeze-and-Excitation (SE) network [28] to generate the final prediction on the emotional state. The scaling vector obtained after the SE network's excitation operation reveals a very important relationship, i.e., the emotional impact from visual attentions, or the influence score IS^t as defined in Section 3.1.

3.3.2 Extracting Eye Features. Eye images $\mathbf{E}^t$ contain information about facial expressions. This work follows the EMO method [61] to extract expression-related features but makes necessary improvements to enhance the emotion recognition accuracy and better suit our application scenarios. Specifically, EMO [61] consists of a feature extractor for eye images and a customized emotional classifier. Since emotion recognition in eye images is not the major pursuit of this work, we only adopt the former (feature extractor based on ResNet-18 backbone

[26], which is denoted as $\mathcal{N}_{eye}$) to extract $\mathbf{f}_e^t \in \mathbb{R}^{128}$, i.e., $\mathbf{f}_e^t = \mathcal{N}_{eye}(\mathbf{E}^t)$ but replace the latter one (the customized classifier) with a binary classifier for neutral/non-neutral predictions. More importantly, we have appended pupil information to $\mathbf{f}_e^t$ before feeding it into the binary classifier. This is inspired by [3], a work showing that statistical eye information such as pupil size can help to improve the emotion recognition accuracy. Denoting the pupil size information as $\mathbf{ps}^t \in \mathbb{R}^2$, we treat $\mathbf{ps}^t$ as expert information, and following prior work [60, 68], we concatenate this expert information $\mathbf{ps}^t$ with $\mathbf{f}_e^t$, which can be written as $\mathbf{f}_{eye}^t = [\mathbf{f}_e^t, \mathbf{ps}^t]$, where the square bracket indicates channel-wise concatenations. Eye features $\mathbf{f}_{eye}^t \in \mathbb{R}^{130}$ encode the expression-related information within eye regions and can be seen as an effective emotional indicator. Note that $\mathcal{N}_{eye}$ will only be applied to eye images when a particular eye attention pattern [8] is identified to save energy, as described in Section 3.4.2. The trigger of the world camera, on the other hand, depends on eye feature $\mathbf{f}_{eye}^t$.

3.3.3 Triggering the World Camera. The high-resolution world camera consumes more energy than the low-resolution eye camera: it would be too energy-intensive to continually capture scene frames. Considering the energy limitations of wearable devices, we have designed a "trigger" mechanism for the world camera. The idea is to skip those emotional-neutral frames (there is no need to analyze the *emotionship* for neutral emotions) and focus on frames with non-neutral emotions. In particular, we design a binary classifier C_{eye} to separate $\mathbf{f}_{eye}^t$ into neutral and non-neutral expressions. If $\mathbf{f}_{eye}^t$ classified as emotionally neutral, the world camera is disabled to save energy. Otherwise, it is triggered to enable the following operations.

3.3.4 Selecting RoI Candidates. The triggered forward-facing world camera records scene images $\mathbf{I}^t$. Using an existing eye-tracking technique [29], we first estimate from $\mathbf{E}^t$ the gaze point $\mathbf{g}^t = (x_g^t, y_g^t)$, where x_g^t and y_g^t refer to the 2D coordinates of this gaze point with respect to scene image $\mathbf{I}^t$. Since $\mathbf{g}^t$ is a 2D point, we still need to find the region of visual perceptions $\mathbf{r}^t = (x_r^t, y_r^t, w_r^t, h_r^t)$. In this work, we use the VinVL model [66] to generate all potential regions in $\mathbf{I}^t$, and then perform filtering to select certain RoI candidates for $\mathbf{r}^t$ from all those regions.

In particular, denote the VinVL model [66] as $\mathcal{N}_{VL}$. Given the scene image $\mathbf{I}^t$, $\mathcal{N}_{VL}$ is able to generate a total of K potential regions $\{\mathbf{R}_1^t, \mathbf{R}_2^t, \ldots, \mathbf{R}_K^t\}$, where $\mathbf{R}_i^t \in \mathbb{R}^4$ represents the i-th candidate. Note that for a RoI $\mathbf{R}_i^t$, its corresponding visual representation (or feature) $\mathbf{f}_{R_i}^t \in \mathbb{R}^{2048}$ and the semantic representation $\mathbf{q}_{R_i}^t$ (e.g., a tag) is already given by $\mathcal{N}_{VL}$. We have designed a filter process to select the ten most suitable regions out of all K regions based on the gaze point $\mathbf{g}^t$. That is, for each candidate $\mathbf{R}_i^t$, we compute the Euclidean distance from its central point to the gaze point $\mathbf{g}^t$. Then we empirically select the top ten regions with the smallest distances, i.e., the ten regions that are closest to the gaze point. These are the most relevant RoIs within the visual attentive region, and are denoted as $\mathbf{R}_c^t = \{\mathbf{R}_{c1}^t, \mathbf{R}_{c2}^t, ..., \mathbf{R}_{c10}^t\}$. After this filtering process, there are still ten RoI candidates, and we need to determine a final visual attention region and also to generate its summary tag. We use two Question Answering (QA) sessions for this purpose.

3.3.5 Determining Visual Attentive Region and Summary Tag. Recall that we have already selected ten candidate regions $\mathbf{R}_c^t = \{\mathbf{R}_{c1}^t, \mathbf{R}_{c2}^t, \ldots, \mathbf{R}_{c10}^t\}$ and the visual feature $\mathbf{f}_{ci}^t \in \mathbb{R}^{2048}$ and its semantic representation $\mathbf{q}_{ci}^t$ of the i-th region $\mathbf{R}_{ci}$ are provided by $\mathcal{N}_{VL}$. To select the actual visual attentive region and generate it summary tags, we perform Visual Question Answering (VQA) [25] and an Image Captioning [2], based on the OSCAR+ vision-language fusion model [66]. Specifically, we denote the visual features of the ten selected regions as $\mathbf{f}_c^t = \{\mathbf{f}_{c1}^t, \mathbf{f}_{c2}^t, \ldots, \mathbf{f}_{c10}^t\}$, the semantic attributes as $\mathbf{q}_c^t = \{\mathbf{q}_{c1}^t, \mathbf{q}_{c2}^t, \ldots, \mathbf{q}_{c10}^t\}$, and the OSCAR+ model [66] as $\mathcal{N}_{OS}$. The VQA session aims to determine the appropriate visual attentive region $\mathbf{r}^t$. First we invoke $\mathcal{N}_{OS}$ to answer the question Q_1 "What object makes people feel happy/surprised/sad/angry/feared/disgusted?" by also inferring to $\mathbf{f}_c^t$ and $\mathbf{q}_c^t$ and obtain an answer $\mathbf{a}^t$ from $\mathcal{N}_{OS}$, which written as

$$\mathbf{a}^t = \mathcal{N}_{OS}(\mathrm{Q}_1, \mathbf{f}_c^t, \mathbf{q}_c^t). \tag{2}$$

Then among the ten attributes $\mathbf{q}_c^t$, we find the one $\mathbf{q}_{cj}^t$ whose word2vec embedding [47] is closest to that of answer $\mathbf{a}^t$ than all other attributes, and $\mathbf{q}_{cj}^t$'s corresponding region $\mathbf{R}_{cj}^t$ is seen as the visual attentive region $\mathbf{r}^t$, i.e., $\mathbf{r}^t = \mathbf{R}_{cj}^t$.

The goal of the Image Captioning (IC) session is to generate an appropriate tag that summarize the semantic attributes of visual perception region. In this session, there is no question to answer, and $\mathcal{N}_{OS}$ only looks at the visual features $\mathbf{f}_c^t$ to generate the tags, which can be expressed as

$$\mathbf{s}^t = \mathcal{N}_{OS}(\mathbf{f}_c^t), \tag{3}$$

where $\mathbf{s}^t$ is the summary tag for visual attentive region.

3.3.6 Determining the Emotional State. The emotional state e^t is obtained through synthesis of eye the features $\mathbf{f}_{eye}^t$ and visual features $\mathbf{f}_c^t = \{\mathbf{f}_{c1}^t, \mathbf{f}_{c2}^t, \ldots, \mathbf{f}_{c10}^t\}$ of the candidate regions. Since $\mathbf{f}_c^t \in \mathbb{R}^{10\times 2048}$ and $\mathbf{f}_{eye}^t \in \mathbb{R}^{130}$, we first employ a Fully Connected (FC) layer to summarize the visual attributes and reduce its dimensionality, i.e., $\mathbf{f}_v^t = \mathcal{FC}_v(\mathbf{f}_c^t)$, where $\mathcal{FC}_v$ denotes the FC layer and $\mathbf{f}_v^t \in \mathbb{R}^{130}$. We concatenate the channels of visual perceptions' feature $\mathbf{f}_v^t$ and eye feature $\mathbf{f}_{eye}^t$ to formulate $\mathbf{f}_{ev}^t = [\mathbf{f}_v^t, \mathbf{f}_{eye}^t]$. This concatenated feature $\mathbf{f}_{ev}^t \in \mathbb{R}^{260}$ contains emotional information from both the eye and scene images, and is fed into a Squeeze-and-Excitation (SE) network [28] $\mathcal{N}_{SE}$ to obtain a scaling vector $\mathbf{u}^t \in \mathbb{R}^{260}$, i.e., $\mathbf{u}^t = \mathcal{N}_{SE}(\mathbf{f}_{ev}^t)$. This scaling vector $\mathbf{u}^t$ is multiplied (channel-wise) with the the concatenated features $\mathbf{f}_{ev}^t$ to obtain feature $\mathbf{f}_{eva}^t \in \mathbb{R}^{260}$, i.e., $\mathbf{f}_{eva}^t = \mathbf{u}^t * \mathbf{f}_{ev}^t$, where $*$ represents channel-wise multiplication. Note that the scaling vector $\mathbf{u}^t$ is learned from the SE gating mechanisms and it reflects the importance degree of each channel in $\mathbf{f}_{ev}^t$. The obtained feature $\mathbf{f}_{eva}^t$ is then input into a soft-max classifier C_{EMO} to generate the final emotion prediction e^t, i.e., $e^t = C_{EMO}(\mathbf{f}_{eva}^t)$.

3.3.7 Computing the Influence Score. The Influence Score indicates the degree of emotional impact from visual perceptions, and it can be computed from the scaling vector $\mathbf{u}^t$ learned from the SE gating mechanism. Recall that $\mathbf{u}^t$ represents the importance of each channel in $\mathbf{f}_{ev}^t$, while $\mathbf{f}_{ev}^t$ consists of eye features $\mathbf{f}_{eye}^t$ concatenated with visual perception's feature $\mathbf{f}_v^t$. We are therefore able to evaluate the importance of the visual perception feature $\mathbf{f}_v^t$ in predicting emotional state, or the Influence Score IS using $\mathbf{u}^t$, which can be written as

$$IS^t = \frac{\sum_{i=1}^{130} \mathbf{u}_i^t}{\sum_i \mathbf{u}_i^t}, \tag{4}$$

where $\mathbf{u}_i^t$ denotes the i-th scalar of $\mathbf{u}^t$ and the first 130 scalars of $\mathbf{u}^t$ corresponds to channels of $\mathbf{f}_v^t$. Using the influence scores in Equation 4, we can determine to which degree an emotional state was affected by the sentimental visions. For instance, if a really small IS^t is computed, we would conclude that the current emotional status is not related to the observed visual perceptions. In contrast, a large IS^t value implies that the current emotion is highly related to the attentive scene regions.

3.3.8 Emotional State in Video Sequence. In the interest of simplicity, we only illustrate how to predict emotion for a certain time step t in the descriptions so far. However, emotions are temporally consistent processes instead of static ones, and therefore we also need to consider how to aggregate emotion predictions in different time steps. For a video clip of T frames, assuming that we have already computed emotion prediction for each frame, i.e. $\{e^1, e^2, ...e^T\}$, we use the most common emotion class e_m as the emotion prediction of this sequence. Formally, the most common emotion class e_m can be computed as

$$e_m = \underset{i}{\operatorname{argmax}} \sum_{j=1}^{T} \mathbb{1}(e^j = i),$$

where $i \in \{0, 1, 2, 3, 4, 5, 6\}$, $\mathbb{1}$ represents the indicator function, $\mathbb{1}(e^j = i) = 1$ if $e^j = i$ and $\mathbb{1}(e^j = i) = 0$ if $e^j \neq i$.

3.4 *EMOShip* System

3.4.1 Hardware Design. The *EMOShip* prototype is a smart eyewear system equipped with two cameras, including one outward-facing world camera and one customized inward-facing eye camera. The outward-facing world camera collects the visual content aligned with the wearer's field of view. We adopt Logitech B525 1080p HD Webcam (1280×960@30fps). The inward-facing eye camera supports continuous eye tracking and eye-related expression feature capture. We use a camera module with a GalaxyCore GC0308 sensor (320×240@30fps) and an IR LED light to illuminate the iris. This hardware module is inspired by the Pupil [3], but we have re-implemented it to suit our needs. The *EMOShip* prototype is equipped with Qualcomm's wearable system-on-module solution, combining Qualcomm Snapdragon XR1 chipset with an eMCP(4GB LPDDR/64GB eMMC). Its typical power consumption is 1 W, which is suitable for battery-powered wearable design.

3.4.2 Software Operation. In *EMOShip*, *EMOShip*-Net performs emotion recognition and *emotionship* analysis. Targeting energy-constrained wearable scenarios, *EMOShip*-Net uses the following energy-efficient workflow.

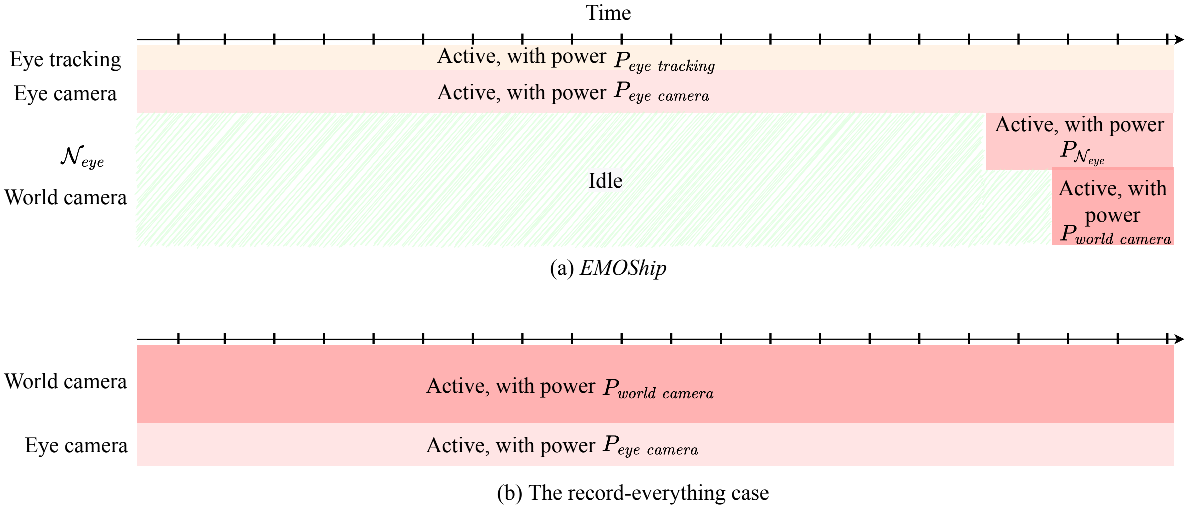
- First, *EMOShip*-Net continuously senses the eye camera to perform eye tracking. To minimize the energy cost of eye tracking, *EMOShip*-Net uses a computationally efficient pupil detection and tracking method [29] to detect potential attention events. Following the work by Chang et al. [8], a potential attention event must satisfy two conditions simultaneously: (1) there is a temporal transition from saccade to smooth pursuit, which suggests a potential visual attention shift and (2) the gaze follows a moving target or fixates on a stationary target. We modify the open-source Pupil software [4] to achieve more accurate eye movement pattern detection that can better satisfy the requirement of our system. Pupil software predicts two eye movements—fixation and non-fixation, based on the degree of visual angles [29]. However, when we deployed it in our system, we needed more eye movements such as saccade and smooth pursuit to detect a potential visual attention event. To address this issue, we follow prior work [8] and leverage the historical gaze trajectory profile to enable a more accurate eye movement detection. This is motivated by the occurrence of smooth pursuit or fixation eye movements when the recent gaze points are located in a constrained spatial region; otherwise, a saccade eye movement occurs. Our goal is the detection of eye movements, so we focus on our method's recall, which is 99.3%. This shows that our eye movement detection method is robust and reliable. The inference time of the eye-tracking method used in *EMOShip*-Net is 115.1 fps, or 8.7 ms/frame.
- Once a potential attention event is detected by the computationally efficient eye-tracking method, *EMOShip*-Net takes the eye images as the input of the light-weight network $\mathcal{N}_{eye}$ to extract eye-related expression-related information and performs neutral vs. non-neutral emotional state classification. $\mathcal{N}_{eye}$ is computationally efficient, which only requires 20.3 ms to perform emotional state classification for each eye image frame.
- Only when a non-neutral emotional state is detected, *EMOShip*-Net activates the high-resolution world camera to sense scene content for semantic analysis. In other words, the high-resolution, energy-intensive scene content capture and processing the pipeline remains off most of the time, avoiding unnecessary data sensing and processing, thereby improving energy efficiency.
- Finally, *EMOShip*-Net leverages a cloud infrastructure to perform computation-intensive semantic attribute feature extraction and eye-scene feature aggregation to support final emotion recognition and *emotionship* analysis, thus offloading energy consumption from the eyewear device.

The energy consumption of the *EMOShip* eyewear device is estimated as follows:

$$E_{EMOShip} = T_{always-on} \times (P_{eye\ camera} + P_{eye\ tracking}) + T_{\mathcal{N}_{eye}} \times P_{\mathcal{N}_{eye}} + T_{captured} \times P_{world\ camera}, \quad (5)$$

[3] https://pupil-labs.com
[4] https://github.com/pupil-labs/pupil/releases

where $T_{always-on}$ is the overall operation time of the *EMOShip* eyewear device, $P_{eye\ camera}$ and $P_{world\ camera}$ are the power consumption of the eye camera and the world camera, respectively, $T_{\mathcal{N}_{eye}}$ is the operation time of the light-weight eye analysis network $\mathcal{N}_{eye}$, and $T_{captured}$ is the operation time of the high-resolution video recording when non-neutral emotional states are detected.



Time
Eye tracking
Active, with power $P_{eye\ tracking}$
Eye camera
Active, with power $P_{eye\ camera}$
$\mathcal{N}_{eye}$
World camera
Idle
Active, with power $P_{\mathcal{N}_{eye}}$
Active, with power $P_{world\ camera}$

(a) *EMOShip*

World camera
Active, with power $P_{world\ camera}$
Eye camera
Active, with power $P_{eye\ camera}$

(b) The record-everything case

Fig. 3. The run-time operation of *EMOShip* (a), as well as that of the recording-everything case (b).

Fig. 3(a) illustrates the run-time operation of *EMOShip* on the eyewear. Physical measurement of the Qualcomm Snapdragon wearable platform shows that $P_{eye\ camera} = 0.07W$, $P_{eye\ tracking} = 0.1W$, $P_{world\ camera} = 1.3W$, and $P_{\mathcal{N}_{eye}} = 1.1W$. Physical measurement during the real-world pilot studies described in Section 5 shows that $\mathcal{N}_{eye}$ and the world camera remain off during 86.8% and 94.6% of the system operation time, respectively. We estimate that a 2.1 Wh battery (similarly to Google Glass Explorer Edition), would allow *EMOShip* to support 5.5 hours of continuous operation without charging.

Had the system continuously recorded (see in Figure 3(b)) with both eye camera and world camera, the overall system energy consumption would have been $E_{always-on} = T_{always-on} \times (P_{eye\ camera} + P_{world\ camera})$, resulting in a battery lifespan of approximately 1.5 hours. Compared with the record-everything case, *EMOShip* improves the system battery lifetime by 3.6×.

4 EVALUATION

This section presents the in-lab experiments to evaluate the performance of *EMOShip*.

4.1 Dataset

To evaluate *EMOShip*, we need the scene images observed by the wearer, the wearer's eye images, and also the emotional states during this observation process. In other words, an eligible dataset should cover both the scene and eye timelines and also contain emotion annotations of the same duration. However, most publicly-available emotion datasets do not satisfy those requirements, since they either lack the scene images, e.g., the MUG dataset [1], or do not provide the facial or eye regions, e.g., the FilmStim dataset [50]. Therefore, we collect and build a new dataset named EMO-Film to suit our needs, which is available online at [5] and detailed below.

[5] https://github.com/MemX-Research/EMOShip

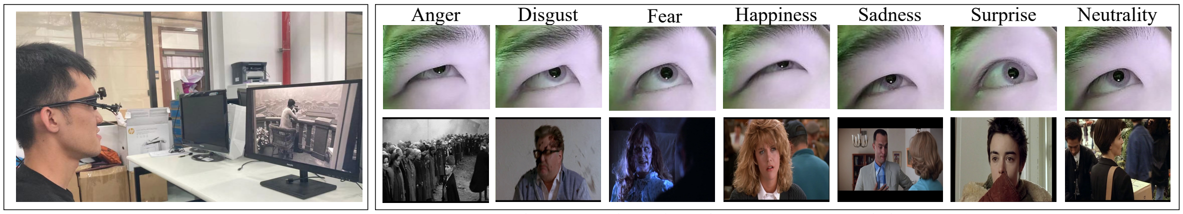
4.1.1 Data Collection. The data of EMO-Film dataset is collected in a controlled laboratory environment. As shown in Fig. 4 (left), participants equipped with *EMOShip* were instructed to watch several emotion-eliciting video clips displayed on a desktop monitor. A total of 20 volunteers attended the data collection of EMO-Film, including 8 females and 12 males.

(1) Video Data Preparation. The video clips were selected from the FilmStim dataset [6] [50] which is a widely used emotion-eliciting video dataset [61]. We first divide all videos of the FilmStim dataset (64 video clips in total) into 7 categories based on the provided sentiment labels, each category corresponding to one emotional class (neutral plus six basic emotions). Then we randomly sample at least one video clip from each category summing up to 6–7 for a participant to watch, which may take approximately 20 minutes to complete. Note that the film clips in FilmStim dataset evoke a broad range of emotional reactions, so this design covers the six basic emotions. We also ensure that each film clip was watched by at least two subjects.

(2) Data Collection Process. During the watching process, we recorded the eye regions of participants using the eye camera. To ensure the video scenes can be captured properly, we pre-adjusted the field of view of the world camera to be aligned with the monitor and recorded the displayed video simultaneously. In this way, we are able to gather the eye/scene data and the emotion ground-truths with aligned timelines, as shown in Fig. 4 (right). This recording session takes approximately 20 minutes per person.

(3) Labeling Process. After all the scheduled movie clips displayed, the participant takes a short break (around 20 minutes) and then is instructed to label their emotional states. This labeling process can take up to 70 minutes (compared with 20 minutes of watching the films) and the generated emotional annotations are arguably accurate since the videos were shown only 20 minutes prior. We develop a labeling tool with a GUI window to facilitate this process. Use of the tool is orally explained to each participant. For each eye/scene image pair, the participant indicates emotional state by clicking on the corresponding button or using a keyboard shortcut. There are a total of seven emotional states to choose from: neutral plus the six basic emotions. We consider only one emotional state per time instant for simplicity. This process is repeated until all eye/scene image pairs have been assigned labels.

The whole data collection process takes approximately four days, and the gathered data and labeling last for approximately 1.5 hours per participant.



(a) Lateral view that a participant is watching video

(b) Examples of eye/scene images and the emotion ground-truths with aligned timelines

Fig. 4. Data collection of EMO-Film: The laboratory setting (left), and the eye and scene images of seven emotional states from the same participant (right).

[6] https://nemo.psp.ucl.ac.be/FilmStim/

4.1.2 Dataset Statistics. EMO-Film dataset is further divided into two sets for the purpose of training/testing, respectively. We split the video data of each subject into 80%/20% for training/testing based on the timestamps. The 80% clips with smaller timestamps (i.e., recorded at an earlier time) are assigned as the training set, and the rest 20% clips as the testing set. The overall percentages of video sequences belong to "anger"/"disgust"/"fear"/"happiness"/ "sadness"/"surprise" are 2.9%/18.2%/20.8%/20.0%/20.8%/17.3%, respectively.

As shown in Table 1, there are a total of 144,145/45,338 eye-scene image pairs in the training/testing set, respectively. Each eye-scene frame pair is properly aligned in timelines, and the frame-level emotion ground-truths are also provided. The resolution for scene images is 1280×960, while that of eye images is 320×240. The distribution of the seven emotion classes is also shown in Table 1. As we can see, "fear" accounts for the most non-neutral emotion events, while "anger" accounts for the fewest. The number of "neutrality" clips is similar to that of "fear". We also apply the data augmentation techniques, including the rotations, flips, and affine transforms, to balance the distribution of different emotional states during the training stage, which can be important to the training of *EMOShip*-Net.

Table 1. Training/testing set distribution.

Emotional States	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutrality
Number of eye-scene image pairs in training set	3,519	21,844	25,000	23,807	24,080	20,895	25,000
Number of eye-scene image pairs in testing set	990	2,843	8,693	4,214	7,068	3,801	17,729

When viewing identical sentimental contents, different participants may demonstrate different emotional reactions. To examine inter-participant variability, we first divide the videos into six sentimental categories excluding neutral, each category corresponding to one emotional class. For each category, we calculate the percentage of video frames for which all subjects demonstrate exactly the same emotional reactions. As shown in Fig. 5, "surprise" can be most easily aroused and shared among people watching the same videos, while there are also a comparatively high proportion of subjects who share the same emotional feelings from viewing content related with "disgust", "fear", and "sadness". "Happiness" and "anger", however, have the lowest probability of commonality accross users, suggesting more personal variation in the perception of such videos.

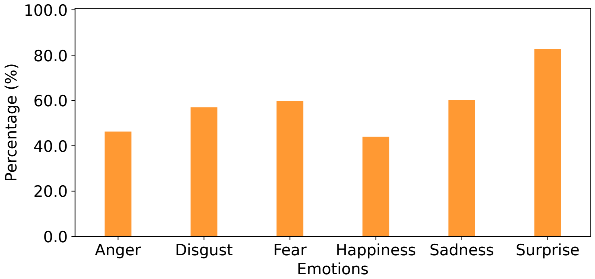


Fig. 5. The percentage of video frames where all subjects demonstrate exactly the same emotional reactions for different emotional classes.

4.2 Experimental Setup

4.2.1 Evaluation Metrics. Emotionship, as defined in Equation 1, captures the emotional states of the users and also describes their potential causes. Since frame-level ground-truth emotion values are provided in our EMO-Film dataset, the evaluation of the former (emotional state prediction) is comparatively straight-forward. Following prior work [65], we adopt the multilabel-based macro-averaging metric to evaluate the performance of emotional state predictions, as defined in Equation (6).

$$B_{macro}(h) = \frac{1}{C} \sum_{j=1}^{C} B(TP_j, FP_j, TN_j, FN_j), \tag{6}$$

where $B(TP_j, FP_j, TN_j, FN_j)$ represents binary classification performance on label j ($B \in \{Accuracy, Precision, Recall\}$). C is the number of emotion classes, in this study, $C = 6$. That is, we only recognize the six non-neutral emotions. TP_j, FP_j, TN_j, and FN_j denote the number of *true positive*, *false positive*, *true negative*, and *false negative* test samples with respect to the j class label, respectively.

However, as *emotionship* itself is a new concept, there is no existing metric that can be used to evaluate its quality, i.e., its accuracy in identifying the causes of emotions. It is also difficult to objectively annotate such potential causes, as they are highly personalized, subjective, and subtle. In this work, several representative samples are visualized to compare the qualities of semantic attributes generated by *EMOShip* and a baseline based on VinVL model [66]. We also plot the variation of Influence Score by scenario to demonstrate whether *EMOShip* has correctly captured the emotional impacts from scene images.

4.2.2 Baselines. To evaluate the performance of emotion recognition, we have selected four works as baselines: 1). the emotion-aware smart glasses EMO [61], 2). EMO+, which is an improved version of EMO, 3). VinVL model [61] that extracts semantic scene features for emotion recognition, and 4). VinVL+ that is modified to focus on the attentive regions of users.

(1) EMO [61] utilizes a deep CNN to recognize emotions from eye images and is closely related with our work. It is used as a primary baseline. Note that we have discarded the classifier in EMO as it requires the construction of an auxiliary face recognition database, which is resource-intensive and brings very limited improvement.

(2) Inspired by prior work [3], we integrate pupil size information with EMO to improve its recognition accuracy. In particular, pupil size is used as a kind of expert information, which is concatenated to the second last Fully Connected (FC) layer of the CNN [60]. This baseline method is denoted as EMO+.

(3) Both EMO and EMO+ predict emotions from the eye images. However, hints on emotional states can also be fetched from scene images, especially from those sentimental visions that are more likely to evoke emotions. To validate this, we devise a third baseline method that predicts emotional states using only the sentimental clues in scene images. In details, we utilize the VinVL model [61] to extract visual features from scene images containing sentimental information. Then, those visual features are fed to a classifier consisting of two layers to obtain the emotion predictions. Regarding the summary tag generation, all the visual features are input into the OSCAR+ model [66] to obtain summary tags. This approach is called VinVL for simplicity.

(4) The visual features of VinVL contain information from all potential Regions of Interest (RoIs). There can be various sentiment clues in those RoIs. However, it is the sentimental information from the user's attentive region that really matters. Therefore, we have set VinVL to consider only those features within the user's attentive region. This method is named VinVL+.

To better understand the causes of emotions, we compare summary tags generated by our *EMOShip* with VinVL+ to provide an intuitive illustration of the qualities.

4.2.3 Training EMOShip-Net. The structure of *EMOShip*-Net is complicated, as it involves several backbone networks with significantly different architectures and design purposes. Instead of end-to-end training, we use an iterative method to train *EMOShip*-Net: each component network is trained individually while freezing the weights of other parts.

The eye network $\mathcal{N}_{eye}$ is used frame-wise and serves as the trigger for the scene camera. It is therefore the first component trained. We generally follow the training procedures in prior work [61] and pre-train $\mathcal{N}_{eye}$ with cross-entropy loss on FER2013 dataset [24] and MUG dataset [1]. Note that on the MUG dataset, the eye regions are cropped out of the facial image, as shown in Fig. 6. The pre-trained $\mathcal{N}_{eye}$ is further fine-tuned the training set of our collected EMO-Film dataset.

Considering the high complexity in visual features model $\mathcal{N}_{VL}$ and the vision-Language model $\mathcal{N}_{OS}$, we directly utilize the pre-trained weights provided by authors of those two models. The Squeeze-and-Excitation model $\mathcal{N}_{SE}$ is trained together with the FC layer $\mathcal{FC}_v$ on EMO-Film dataset. We use the Adam [31] optimizer with an initial learning rate of 0.001 and the batch size is set to 512. The whole training process lasts for a total of 500 epochs.

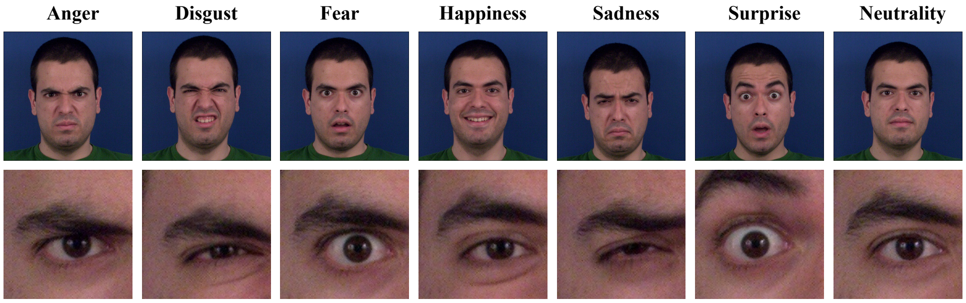


Fig. 6. Seven emotional expressions of the original MUG facial expression examples (top row), our fine-tuning single-eye-area data cropped from MUG (bottom row).

4.3 Results

4.3.1 Emotion Recognition. We present the performance of emotion recognition from the following two aspects.

(1) Binary Emotion Classification. In our system, the neutral/non-neutral classification results from $T_{\mathcal{N}_{eye}}$ serve as a trigger to capture emotional moments, and the accuracy of this binary classification can directly affect the performance of the whole system. Therefore, we first examine the quality of binary classification model EMO+ to determine whether the triggering system is reliable. As shown in Table 2, EMO+ significantly outperforms the baseline EMO model and achieves 80.7% precision, 79.0% recall, and 80.4% accuracy on this binary classification task. This demonstrates the value of adding pupil information in EMO models. The high accuracy achieved by EMO+ also indicates that *EMOShip*-Net is sensitive to emotional moments.

Table 2. Performance comparison of binary emotion classification.

Method	Precision	Recall	Accuracy
EMO+	80.7%	79.0%	80.4%
EMO	78.1%	74.6%	76.9%

(2) Multiple Emotion Classification. Table 3 demonstrates the emotion recognition performance of the four baseline methods and *EMOShip*-Net on EMO-Film dataset.

EMO [61] significantly outperforms VinVL [66] in terms of precision, recall, and accuracy. This is expected, as the emotional clues within eye images are more generally straightforward compared with the indirect and subtle sentimental clues in scene images. EMO+, the improvement version of EMO, has superior performance to EMO, indicating the value of integrating pupil size information. The performance of VinVL+ also surpasses that of VinVL, which illustrates the importance of involving user attention. However, VinVL+ still cannot outperform EMO and EMO+, indicating the importance of expression-related features.

Different from those baselines, *EMOShip*-Net fuses emotional evidence of both scene and eye images to achieve more comprehensive and accurate emotion predictions. Notably, EMOShip-Net significantly outperforms the best baseline EMO+ by 5.3% precision, 5.8% recall, and 6.0% accuracy. This reveals the importance of inferring from both facial expressions and visual perceptions, and indicates the superiority of *EMOShip*-Net in determining emotional states.

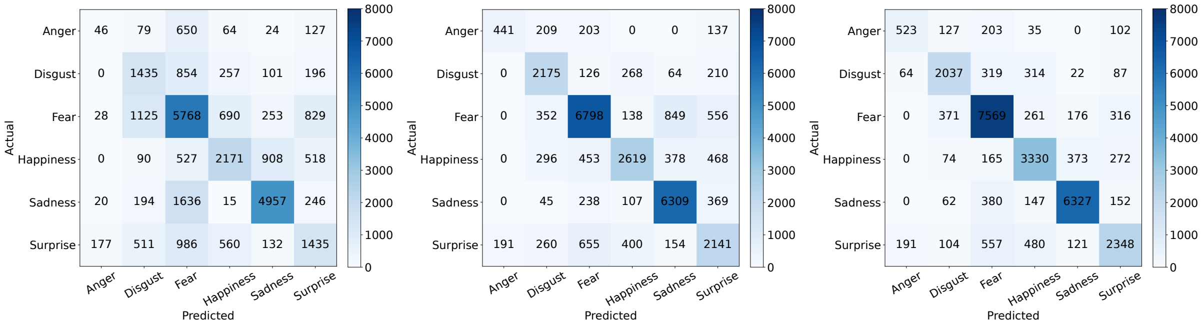
Table 3. Performance comparison of multiple emotion classification for the proposed method and the baseline methods.

Method	Precision	Recall	Accuracy
EMOShip-Net (Ours)	76.3%	73.6%	80.2%
EMO+	71.0%	67.8%	74.2%
EMO [61]	65.9%	67.0%	69.4%
VinVL+	48.8%	46.8%	57.3%
VinVL [66]	42.6%	44.3%	55.5%

We have plotted the confusion matrices of different methods. Figs. 7a, 7b, and 7c demonstrate that *EMOShip*-Net achieves a better recognition rate for most emotions, demonstrating its superior generalization ability. We also observe that *EMOShip* performs slightly worse on “disgust” than EMO+. That is because EMO+ determines emotional states exclusively based on eye images, while *EMOShip* takes both the visual region and eye images into consideration. This may undermine accuracy when *EMOShip* receives strong misleading signals from visual attentive regions. For example, when scene images containing emotionally negative material are captured, it can be challenging for *EMOShip* to determine which kind of negative emotions (such as “disgust” or “fear”) should be related to this visual information since they may all occur in response to negative scenes. As shown in Figs. 7a and 7b, the VinVL+ method, which only utilizes visual information, generally delivers lower classification rates on negative emotions such as “disgust” and “anger” than EMO+, while its recognition accuracy on other classes, such as “happiness” and “sadness”, are relatively similar. In conclusion, associations between negative sentimental visions and negative emotions can be challenging to establish.

Fig. 8 shows an exemplary case to provide further intuition. It presents successive scene/eye image sequence and corresponding emotion predictions within an approximate six-second clip generating “fear” emotions. Both VinVL+ and EMO+ baseline have produced inconsistent emotion predictions during this clip, while our method has successfully predicted fear as the result of viewing all those frames. This demonstrates that our method produces more temporal-consistent emotional predictions, thanks to the fusion of emotional evidence from both visual scenes and eye regions.

4.3.2 Understanding of Potential Cause of Emotions. Understanding the cause of emotions can be too subtle and subjective to be quantitatively evaluated, especially when we are aiming to compute the Influence Score *IS*, i.e., the intensity of emotional response to visual perceptions. Despite those challenges, *EMOShip* is the first eyewear system to reveal the semantic attributes of visual attention and to associate those attributes with varying



(a) VinVL+ (b) EMO+ (c) *EMOShip*-Net (Ours)

Fig. 7. Confusion matrix of individual emotional moments when using the two baseline methods and the proposed method.

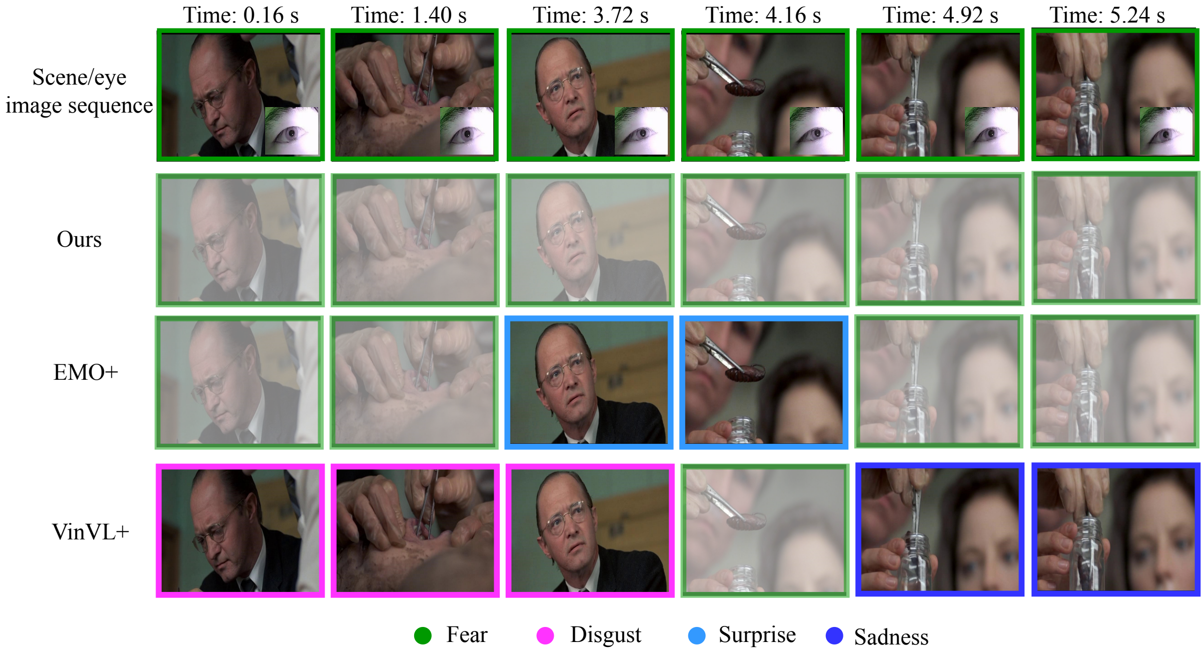


Fig. 8. An example of emotion recognition comparison between the proposed *EMOShip*-Net and the two baseline methods. Colored rectangles highlight the true emotions and the emotions predicted by these methods.

emotional states. We analyze such behaviors using several examples. In particular, we adopt cases of different emotions to present the qualities of the summary tags of visual attentive regions, and then we plot the Influence Score from real video clips to examine their temporal patterns.

In Fig. 9, we provide examples of the semantic summary tags generated by our method and the VinVL baseline. A general trend can be discovered that the summary tag of our *EMOShip* has better captured the sentimental clues in those scenarios than has the VinVL baseline. For example, in the "fear" case, the summary tag of *EMOShip* contains emotion-indicating keywords such as “screaming”, which is highly relevant with negative emotions like “fear” and is clear evidence of awareness of the sentimental visions. In contrast, VinVL displays neutral descriptions and uses words like “talking” to depict this scene, which are less sentimentally accurate. Similar observations can be made on other emotions where *EMOShip* uses more emotional indicators such as “dirty room”, “screaming face”, “dark room”, etc. These differences are not difficult to understand. The visual features used in the VinVL baseline method are not filtered and consist of visual information from non-attentive regions. Such irrelevant information can confuse the language model and can lead to less appropriate summary tags like the sentimentally neutral words. In contrast, *EMOShip* uses the selected visual features (see Section 3.3.4 for details) that are highly relevant to the visual attentive region, thus generating summary tags that are more relevant to the visual attentive region and more likely to contain sentimentally non-neutral meanings.

The tags of our eyewear device are generally more semantically accurate than VinVL. Using “fear” as an example, *EMOShip* correctly depicts the scenario, i.e., “A young girl is screaming while sitting on a bench”, while VinVL’s description is less appropriate, i.e., “A young girl is talking on a cellphone”. This semantic accurateness is also an advantage of *EMOShip*.

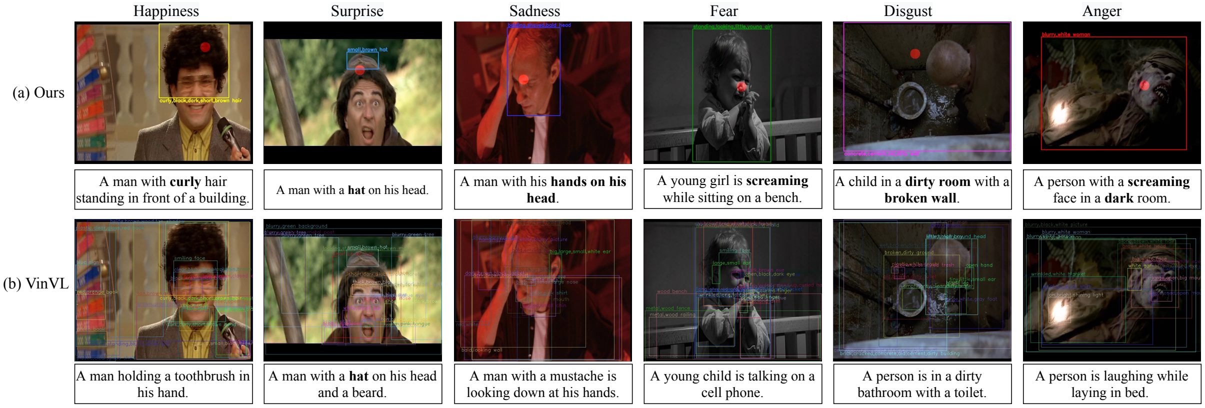

Happiness
Surprise
Sadness
Fear
Disgust
Anger
(a) Ours
A man with **curly** hair standing in front of a building.
A man with a **hat** on his head.
A man with his **hands on his head**.
A young girl is **screaming** while sitting on a bench.
A child in a **dirty room** with a **broken wall**.
A person with a **screaming** face in a **dark** room.
(b) VinVL
A man holding a toothbrush in his hand.
A man with a **hat** on his head and a beard.
A man with a mustache is looking down at his hands.
A young child is talking on a cell phone.
A person is in a dirty bathroom with a toilet.
A person is laughing while laying in bed.

Fig. 9. Examples of the semantic summary tags generated by our method (a) and the VinVL baseline (b). The summary tag is shown at the bottom of each frame. The red circle indicates the gaze point. The emotional words are highlighted in bold.

Next, we investigate how different emotional states can be associated with scene features through the use of Influence Score *IS*. Fig. 10 shows the normalized average *IS* of six non-neutral emotional categories. We can observe that the emotion “sadness” exhibits the highest the *IS* value. This indicates that emotion “sadness” is generally more tightly associated with our visual perceptions than others. Also, emotion “surprise” presents the lowest *IS* score and is therefore considered to be less related with scene features than all other emotions.

4.3.3 Generalization Ability. We also examine the generalization ability of *EMOShip*-Net on unseen users. Specifically, 5 new participants out of the EMO-Film dataset were recruited, and we follow identical data collection procedures as in Section 4.1.1 to produce an evaluation data set that is strictly subject-independent with the EMO-Film dataset used to train our models. This new evaluation set is approximate 105 minutes in length. We evaluate the in-lab emotion recognition performance of *EMOShip*-Net on this newly-collected unseen dataset. In particular, we compare the performance of *EMOShip*-Net on this new test set with that of the EMO and EMO+ baseline methods (the two most out-standing baseline methods). The results are shown in Table 4. We can see that *EMOShip*-Net has superior performance than EMO and EMO+. This indicates that *EMOShip*-Net can generalize well to unseen subjects.

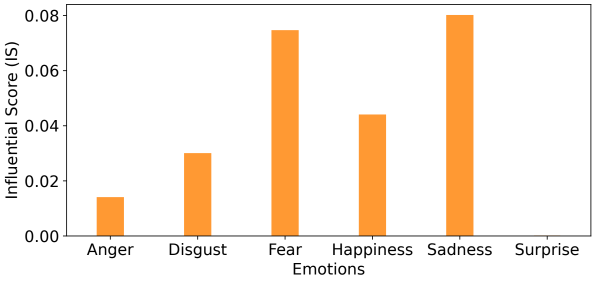


Fig. 10. Degree of emotional impacts from visual perceptions.

Table 4. Performance comparison of multiple emotion classification for new/unseen users.

Method	Precision	Recall	Accuracy
EMOShip-Net (Ours)	65.9%	76.6%	78.5%
EMO+	62.8%	65.4%	70.0%
EMO	61.2%	60.3%	61.4%

We further examine the performance regarding F1 scores of different methods on two test sets. The performance of one method differs depending on test set, because one may be more challenging than the other. We compute F1 scores of different methods on the original/new test set along with the drop rates. As shown in Table 5, all methods perform worse on the new test set: we may reasonably assume that it is more challenging than the first. Nonetheless, the performance of *EMOShip*-Net degrades the least on this new set, suggesting the importance of exploiting *emotionship*.

Table 5. Performance comparisons regarding F1 score on subject dependent/independent test sets, respectively.

Method	F1 Score		Drop Rate
	Original Test Set (subject-dependent)	New Test Set (subject-independent)	
EMOShip-Net (Ours)	74.9%	70.8%	5.4%
EMO+	69.4%	64.1%	7.6%
EMO	66.4%	60.7%	8.6%

5 PILOT STUDY

In additional to in-lab experiments, we also performed a approximate three-week in-field pilot study to evaluate the performance of *EMOShip* under realistic scenarios. In this section, we present two real-life applications of *EMOShip*, demonstrate its usability outside the laboratory, and describe its limitations and promising directions for future work.

5.1 Applications

The most significant advantage of *EMOShip* is that it captures *emotionship* instead of emotions. Compared with other emotional-aware glasses like EMO [61], our *EMOShip* not only predicts emotions at higher accuracy but also provides intuitive explanations on the potential causes of those emotional states. This awareness of *emotionship* opens the door to new applications. Multiple rounds of user interviews lead to two applications: Emotionship Self-Reflection and Emotionship Life-Logging.

In psychology, the term "self-reflection" refers to the process of analyzing past behaviors to achieve better efficiency in the future [16, 23]. Self-reflection is indispensable, especially for people affected by negative emotions. As indicated in relevant studies [17], negative emotions can lead to mental well-being issues. To maintain mental health, we need to self-reflect on negative emotional moments, and we also need to find what evokes those emotions so exposure to those causes can be minimized. This scenario is appropriate for *EMOShip*, which has the ability to record emotional moments, retrieve negative emotional moments, and discover their causes. We name this application Emotionship Self-Reflection.

Life-logging is usually considered to be digital self-tracking or recording of everyday life [52]; it is already a popular application. Current life-logging applications commonly record scenes with commercial glasses like GoPro and Google Clip. It is also difficult to categorize such recordings into different emotional categories, since those eyewear devices lack emotion awareness. Manually classifying emotional moments is extremely time-consuming and tedious, and the user may not be able to recall the extracted emotional activities. Therefore, we integrated *EMOShip* and life-logging to produce a new application called Emotionship Life-Logging, which can automatically detect emotional moments, record them, and document their causes. Emotionship Life-Logging also enables various interesting and promising down-stream tasks such as retrieving and classifying emotional moments [64].

5.2 Procedure of Pilot Study

In-field pilot studies are performed for the two applications described above. A total of 20 volunteers, including 14/6 males/females aged between 23 to 40, were recruited to participate in pilot studies. The research goal of understanding the potential causes of emotions was indicated to all participants before the pilot studies. Volunteers were also informed that their daily activities would be recorded for research purposes.

During this in-field pilot study, participants were introduced to wear *EMOShip* whenever practical to maximize coverage of their day-to-day lives. *EMOShip* automatically recorded emotional moments along with their potential (attention-related) visual causes. The complete scene videos taken by the world camera were also saved for reference and are referred to as the baseline video.

The pilot studies lasted for approximately three weeks. At the end of the study, volunteers were asked to assist in evaluating the value of *EMOShip* for understanding daily emotions and their causes. In particular, we required participants to 1) watch the emotional moments captured by *EMOShip* and mark those clips they believed to have correctly reflected their emotional states and 2) retrieve from the baseline video emotional moments that *EMOShip* failed to capture. In addition, the participants were asked to complete a questionnaire survey of their opinions on the usability and value of the two emotionship applications.

5.3 Performance of *EMOShip* in Pilot Study

5.3.1 Quantitative Evaluations. Table 6 summarizes the system performance of *EMOShip*. The participants generated a total of 530.7 minutes of baseline video and 33.8 minutes of 212 emotional video clips. Compared with the overall operation time, $T_{always-on} = 530.7\,min$, the operation time reduction for the eye features extraction and high-resolution video capturing are 84.2% ($\frac{T_{always-on}-T_{N_{eye}}}{T_{always-on}}$) and 93.6% ($\frac{T_{always-on}-T_{capture}}{T_{always-on}}$), respectively. That is consistent with the short-term property of non-neutral emotions. As indicated in relevant research [56], non-neutral emotions are typically aroused by sudden emotional stimuli, and are short-term mental processes that can vanish in a few seconds. In other words, non-neutral emotions are much rarer than neutral ones in daily life. We inspect P6 for a detailed understanding. One of the scenarios of P6 is watching a basketball game lasting for

Table 6. Overall performance of the in-field pilot study. $T_{always-on}$ is the overall operation time of *EMOShip*. $T_{N_{eye}}$ and $T_{capture}$ are the operation time of eye tracking and the high-resolution video recording, respectively. EM means the emotional moments.

Participant	$T_{always-on}$ (minute)	$T_{N_{eye}}$ (minute)	$T_{capture}$ (minute)	# of Distinct EM	# of True EM	# of False EM	# of Missed EM	Precision	Recall
P1	24.8	4.7	2.0	3	17	3	2	85.0%	89.5%
P2	28.5	5.4	3.6	2	10	2	1	83.3%	90.9%
P3	42.5	2.7	2.0	1	11	1	2	91.7%	84.6%
P4	55.6	7.0	2.1	3	19	6	4	76.0%	82.6%
P5	32.7	3.7	2.7	2	13	4	3	76.5%	81.3%
P6	73.2	8.9	2.2	3	23	7	4	76.7%	85.2%
P7	44.7	2.4	1.2	1	8	1	0	88.9%	100.0%
P8	17.9	1.9	1.1	1	8	2	3	80.0%	72.7%
P9	26.9	6.9	1.2	2	9	3	3	75.0%	75.0%
P10	16.0	4.2	1.5	1	7	1	0	87.5%	100.0%
P11	17.3	4.6	1.4	4	13	2	3	86.7%	81.3%
P12	10.9	2.7	0.6	3	3	0	2	100.0%	60.0%
P13	32.7	4.8	2.0	4	13	1	5	92.9%	72.2%
P14	14.5	2.4	1.3	2	9	1	0	90.0%	100.0%
P15	12.3	2.9	1.5	3	8	1	2	88.9%	80.0%
P16	14.2	3.2	1.4	2	8	3	1	72.7%	88.9%
P17	11.8	3.1	1.2	1	8	0	1	100.0%	88.9%
P18	24.2	4.0	2.2	2	13	2	4	86.7%	76.5%
P19	14.3	3.2	0.8	2	4	0	2	100.0%	66.7%
P20	15.7	5.1	1.8	3	8	4	1	66.7%	88.9%
Mean								**82.8%**	**83.1%**

around 12 minutes, within which our system has detected 0.4 minutes of non-neutral Emotional Moments (EM). Those EMs occurred exactly when the wearer sees two scoring shots, each one lasting for around 0.2 minute. Given a 30 fps sampling rate, the 2 EMs contain approximate 720 image frames (2×30×0.2×60). Apart from those moments, P6 remains emotionally neutral. *EMOShip* correctly captures those non-neutral emotional moments.

Based on emotional moments marked by users at the end of the pilot study, we are able to evaluate the performance of *EMOShip* in practice. Generally, users pay attention to how many emotional moments are correctly recorded by *EMOShip*, and how many emotional moments are missed or incorrectly recorded. We use precision $\frac{Number\ of\ True\ EM}{Number\ of\ True\ EM+Number\ of\ False\ EM}$ to indicate the former, and recall $\frac{Number\ of\ True\ EM}{Number\ of\ True\ EM+Number\ of\ Missed\ EM}$ to indicate the latter. Results show that, *EMOShip* delivers 82.8% precision and 83.1% recall on average, which means that *EMOShip* can accurately capture personal emotional moments, and most of the emotional moments can be captured by *EMOShip*.

We also plot a confusion matrix for the pilot studies to provide a more intuitive understanding. As shown in Fig. 11, *EMOShip* has high emotional category classification accuracy. Positive emotions [10] (171 of "happiness" and "surprise") are much more frequent than negative ones (53 of "sadness", "anger", "disgust" and "fear"), indicating that positive emotions are the dominate emotional states in the daily lives of our pilot study participants.

5.3.2 Emotionship Analysis.

We first summarize the emotional states of all participants and then give a concrete example to provide further insights.

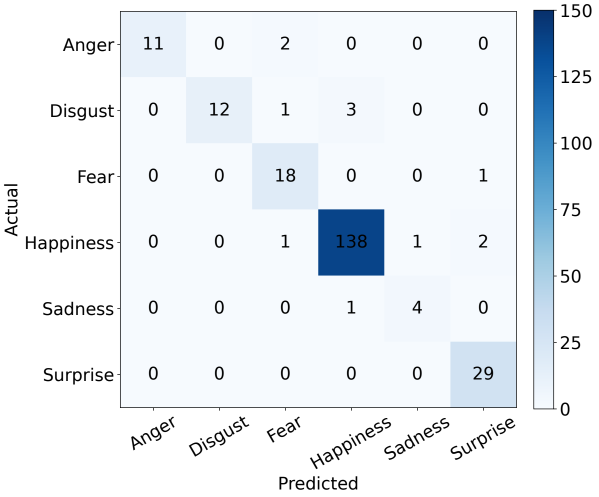


Fig. 11. Confusion matrix of individual emotional moments when using *EMOShip* in pilot studies.

(1) Emotional States Summary. To give participants an overall understanding of their past emotional states, we briefly summarize the past emotional states for each user by roughly categorizing the six basic non-neutral emotional states as *positive* and *negative*. Intuitively, we categorize "happiness" and "surprise" as positive emotional states, while the remaining four are categorized as negative. For each user, we use Pr and Nr to denote the proportion of positive emotions and negative emotions, respectively. For a certain time window, we can suggest two rough emotional patterns as follows:

- Type I: $Pr > Nr$, indicating that the overall emotional state of a user lean towards positive.
- Type II: $Pr \leq Nr$, suggesting that a user is more frequently occupied by negative emotions.

As shown in Fig. 12, we can observe that 17 out of 20 users belong to Type I, while 3 users fall into Type II (P8, P9, and P11), indicating that positive emotions are the dominating emotional states during the pilot studies.

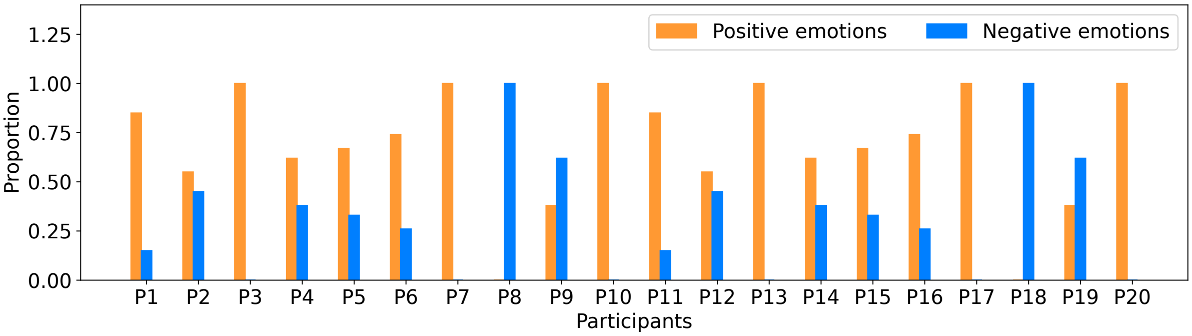


Fig. 12. Profile of emotional states for all 20 participants.

(2) An Exemplary Case. The following example provides further intuition on how *EMOShip* works. Fig. 13 shows the temporally consistent emotional states for a participant (P6). P6 was selected due to being the most

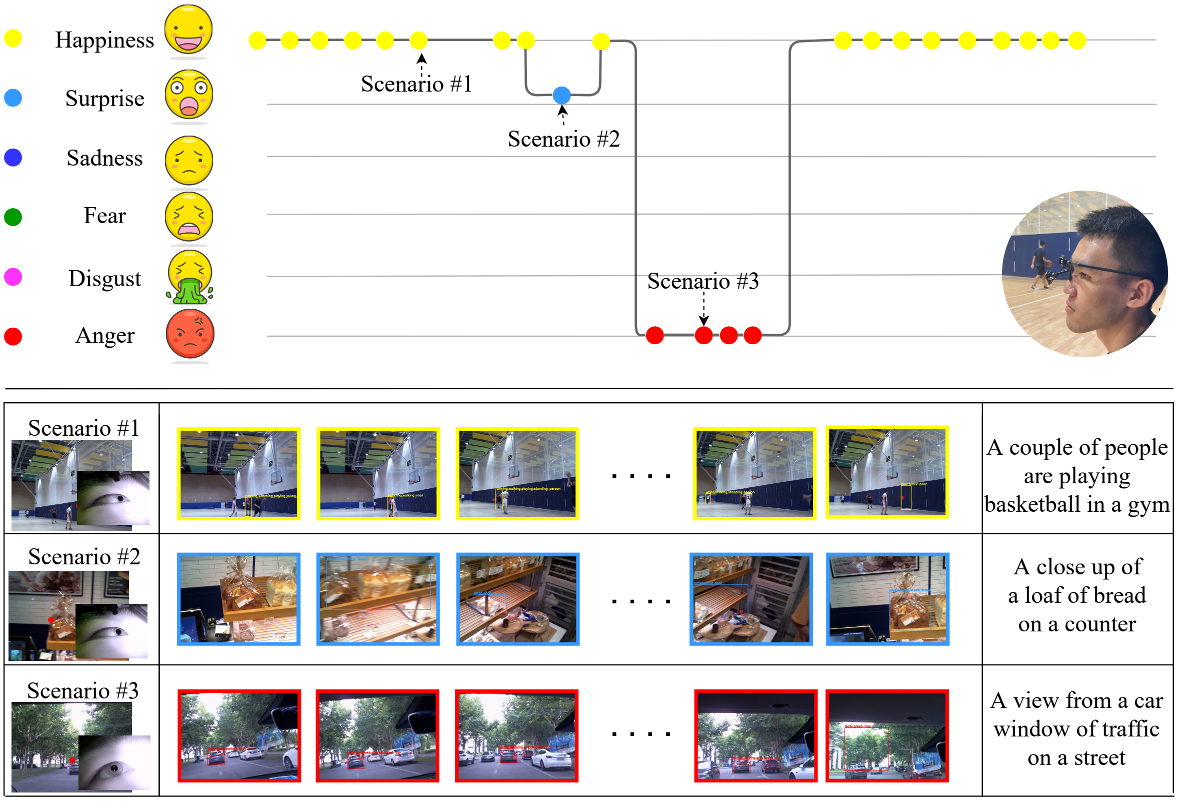


Fig. 13. Time series emotional states for a participant (P6). Emojis are taken from [42].

active user during the pilot study, leading to 23 emotional clips with a duration of 8.9 minutes covering most of the emotional categories, which provides us with good opportunities to explore the insight of *EMOShip*.

As can be seen from Fig. 13, during the whole timeline, the major emotional state is "happiness". This is not surprising, as we can examine the corresponding scenario, i.e., "Scenarios #1" in the figure, and we can see from the summary tag that "A couple of people are playing basketball in a gym". This tag, along with the scenario image, indicates that this user is actually enjoying watching a basketball game and is quite likely to be happy. When it comes to the "surprise" case of "Scenarios #2", we can derive from the tag and also the attentive region that P6 is surprised to see a close-up of a loaf of bread. As for the "anger" of "Scenarios #3", we can immediately learn from the summary tag and attentive region that P6 is driving a car and feels angry due to the traffic. In a similar way, this participant can easily access all the emotional moments that are valuable and personalized. If this user would like to perform emotionship self-reflection, those anger moments can be retrieved to discover what led to the emotion, e.g., traffic, making it easier to avoid such situations in the future. In summary, *emotionship* has application in the valuable application of self-reflection.

5.3.3 Feedback from Participants. We also used a questionnaire to ask study participants about their opinions of the two applications, their wearing experience, and request ideas for improvements. In summary, 16 out of 20 participants provided positive feedback on Emotionship Self-reflection, while 15 out of 20 people saw value in Emotionship Life-logging.

We selected several illustrative comments from participants and quoted them as follows.

One participant remarked: *"From my experience, EMOShip has allowed me to recognize and understand my emotions in a major meeting, which was quite profound to me. When I rewatched the video clips and emotions recorded by EMOShip, I realized that I appeared to be very negative during the meeting, and the meeting was also quite heavy. If I had noticed these issues then, I believe I would have been able to readjust myself to encourage the participation of the team and have a more productive meeting. So I think I will use EMOShip in more meetings and social events. In the long run, it would be significantly beneficial for me to understand and manage my emotions by utilizing EMOShip to analyze my emotions and record my emotional moments. "—P1*

Similarly, another participant appreciated the application of *EMOShip* to long-term mood perception and management, as figured out by this volunteer: *"EMOShip shows that I have two significantly different states of mind when driving or walking to work. When I commute on foot, the emotions appear to be more positive and I tend to feel happy more frequently. My driving emotions, on the other hand, often seem to be negative, such as fear and anger. ...... I may feel negative or get road rage encountering rule-breaking behaviours such as slow left-lane driving or unsafe lane changes. In addition, with the help of EMOShip I also noticed that I seem to be overly cheerful during business meetings, which may leave an unintended impression of me being unprofessional or unreliable. EMOShip unveils the importance of facial expression management to me. I need to be more aware of my social environment whether I should be more happy or serious. "—P2*

The third user stated that *EMOShip* can significantly ease the logging of emotional moments, which can be of importance: *"EMOShip can assist me to record some interesting or important moments and my emotions at that time, both of which are crucial for me to get these moments rapidly reviewed. .... Reviewing the meeting materials that are important to me by watching the videos EMOShip recorded can save me a great amount of time. Plus, my emotions may also shift during interesting moments in life. For example, EMOShip records intense game sessions and sensational videos when I feel happy or sad. It would have been very inconvenient for me to record them manually clip by clip while playing games or watching videos, whereas EMOShip can easily record them for me to review or share quickly afterwards. "—P6*

On the other hand, there is also a volunteer who disregarded the importance of recording emotional moments, and we quote his feedback below: *"I used EMOShip while playing cards. Since this is a highly enjoyable entertainment, there was little change in my recorded emotion types. Moreover, I probably didn't pay much attention to the changes in my emotions. "—P7*

5.4 Limitations and Future Works

We have demonstrated the technical capabilities of *EMOShip* to recognize emotion states and understand their causes. However, we also observe several limitations from its applications to real-world scenarios and from users' feedback. In this section, we briefly discuss some potential future works that will further improve *EMOShip* system.

5.4.1 Personalized Emotional Management. Although most users have provided positive feedback on *EMOShip*, a consensus is that they would like to also receive suggestions on how to reduce the occurrences of negative emotional moments. Since different people have different situations and hence require personalized service, we are planning to integrate a long-term emotion tracking, emotional management, and regulation system into *EMOShip*, which can be personalized to suggest how to avoid causes of negative emotions.

5.4.2 Privacy Concern and Privacy Protection. Although participants are interested in perceiving their emotional states, some participants are uncomfortable with exposing their personal affective information to third parties. Future system design should carefully consider how to address privacy concerns. Another common feedback from users is that they are worried about the disclosing of their emotional information, especially to malicious third parties. Therefore, we set plans on enhancing the privacy protection of using *EMOShip* and also on ensuring the safety of recorded personal data, from both software and hardware sides.

5.4.3 Multi-Modality in Emotional Causes. *EMOShip* focuses on visual stimuli as stimulus for emotions. However, the other senses are also important. For example, the auditory perception, like a sharp, annoying sound, can also affect emotional states. Fusing emotionally relevant features from multi-modal data remains a challenging topic to address in the future.

6 CONCLUSIONS

This paper has proposed and defined the *emotionship* analysis problem for eyewear devices. It has described a deep neural network, called *EMOShip*-Net, that predicts semantic attributes from scene images and can synthesize emotional clues from both eye and scene images and is aware of each feature's importance. Based on *EMOShip*-Net, we present *EMOShip*, the first-ever intelligent eyewear system capable of *emotionship* analysis. Experimental results on the FilmStim dataset of 20 participants demonstrate that *EMOShip* captures emotional moments with 80.2% accuracy, significantly outperforming baseline methods. We also demonstrate that *EMOShip* can provide a valuable understanding of the causes of emotions. We developed two applications using *EMOShip*: emotionship self-reflection and emotionship life-logging. A 20-user in-field pilot study demonstrated that most participants think *EMOShip* helps them reflect on and understand their emotions they usually ignore. Most participants indicated that *EMOShip* was of value to them. *EMOShip* is the first embodiment of an *emotionship*-aware eyewear system. It is relevant to the coming age of intelligent wearable systems and brings us closer to answering the question, "Will smart glasses dream of sentiment visions?"

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61932007 and 62090025 and in part by the National Science Foundation of the United States under grant CNS-2008151.

REFERENCES

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. 2010. The MUG facial expression database. In *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, 1–4.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5561–5570.

[3] Claudio Aracena, Sebastián Basterrech, Václav Snáel, and Juan Velásquez. 2015. Neural networks for emotion recognition based on eye tracking data. In *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2632–2637.

[4] Lisa Aziz-Zadeh and Antonio Damasio. 2008. Embodied semantics for actions: Findings from functional brain imaging. *Journal of Physiology-Paris* 102, 1-3 (2008), 35–39.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of ICLR*. http://arxiv.org/abs/1409.0473

[6] Mark Blum, Alex Pentland, and Gerhard Troster. 2006. Insense: Interest-based life logging. *IEEE MultiMedia* 13, 4 (2006), 40–48.

[7] Victor Campos, Brendan Jou, and Xavier Giro-i Nieto. 2017. From pixels to sentiment: Fine-tuning CNNs for visual sentiment prediction. *Image and Vision Computing* 65 (2017), 15–22.

[8] Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P. Dick, Tun Lu, Ning Gu, and Li Shang. 2021. MemX: An Attention-Aware Smart Eyewear System for Personalized Moment Auto-Capture. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 56 (June 2021), 23 pages. https://doi.org/10.1145/3463509

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations. (2019).

[10] Stephen D Christman and Michelle D Hackworth. 1993. Equivalent perceptual asymmetries for free viewing of positive and negative emotional expressions in chimeric faces. *Neuropsychologia* 31, 6 (1993), 621–624.

[11] Maarten Coëgnarts and Peter Kravanja. 2016. Perceiving causality in character perception: A metaphorical study of causation in film. *Metaphor and Symbol* 31, 2 (2016), 91–107.

[12] Rebecca J Compton. 2003. The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and cognitive neuroscience reviews* 2, 2 (2003), 115–129.

[13] Jean Costa, Alexander T Adams, Malte F Jung, François Guimbretière, and Tanzeem Choudhury. 2016. EmotionCheck: leveraging bodily signals and false feedback to regulate our emotions. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 758–769.

[14] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. 2001. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine* 18, 1 (2001), 32–80.

[15] Tim Dalgleish. 2004. The emotional brain. *Nature Reviews Neuroscience* 5, 7 (2004), 583–589.

[16] Marilyn Wood Daudelin. 1996. Learning from experience through reflection. *Organizational dynamics* 24, 3 (1996), 36–48.

[17] Elena Di Lascio. 2018. Emotion-Aware Systems for Promoting Human Well-being. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 529–534.

[18] Elena Di Lascio, Shkurta Gashi, and Silvia Santini. 2018. Unobtrusive Assessment of Students' Emotional Engagement during Lectures Using Electrodermal Activity Sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 103 (Sept. 2018), 21 pages. https://doi.org/10.1145/3264913

[19] Samira Ebrahimi Kahou, Vincent Michalski, Kishore Konda, Roland Memisevic, and Christopher Pal. 2015. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 467–474.

[20] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.

[21] Kurara Fukumoto, Tsutomu Terada, and Masahiko Tsukamoto. 2013. A smile/laughter recognition mechanism for smile-based life logging. In *Proceedings of the 4th Augmented Human International Conference*. 213–220.

[22] Jose Maria Garcia-Garcia, Victor MR Penichet, and Maria D Lozano. 2017. Emotion detection: a technology review. In *Proceedings of the XVIII international conference on human computer interaction*. 1–8.

[23] Surjya Ghosh, Bivas Mitra, and Pradipta De. 2020. Towards Improving Emotion Self-Report Collection Using Self-Reflection. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[24] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. 2013. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*. Springer, 117–124.

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6904–6913.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[27] Steven Hickson, Nick Dufour, Avneesh Sud, Vivek Kwatra, and Irfan Essa. 2019. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1626–1635.

[28] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7132–7141.

[29] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 1151–1160.

[30] Bo-Kyeong Kim, Hwaran Lee, Jihyeon Roh, and Soo-Young Lee. 2015. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 427–434.

[31] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[32] Gil Levi and Tal Hassner. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 503–510.

[33] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

[34] Shan Li and Weihong Deng. 2020. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing* (2020).

[35] Shan Li, Weihong Deng, and JunPing Du. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2852–2861.

[36] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

[37] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).

[38] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of EMNLP*. 1412–1421. https://doi.org/10.18653/v1/d15-1166

[39] Pingchuan Ma, Yujiang Wang, Jie Shen, Stavros Petridis, and Maja Pantic. 2021. Lip-reading with densely connected temporal convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2857–2866.

[40] Jana Machajdik and Allan Hanbury. 2010. Affective image classification using features inspired by psychology and art theory. In *Proceedings of the 18th ACM international conference on Multimedia*. 83–92.

[41] Katsutoshi Masai, Yuta Sugiura, Katsuhiro Suzuki, Sho Shimamura, Kai Kunze, Masa Ogata, Masahiko Inami, and Maki Sugimoto. 2015. AffectiveWear: towards recognizing affect in real life. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*. 357–360.

[42] Vincent Le Moign. 2017. Streamline Emoji, Free Icons from the Streamline Icons Pack. https://streamlineicons.com.

[43] Arne Öhman, Anders Flykt, and Francisco Esteves. 2001. Emotion drives attention: detecting the snake in the grass. *Journal of experimental psychology: general* 130, 3 (2001), 466.

[44] Hadas Okon-Singer, Jan Mehnert, Jana Hoyer, Lydia Hellrung, Herma Lina Schaare, Juergen Dukart, and Arno Villringer. 2014. Neural control of vascular reactions: impact of emotion and attention. *Journal of Neuroscience* 34, 12 (2014), 4251–4259.

[45] Maja Pantic and Leon JM Rothkrantz. 2003. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE* 91, 9 (2003), 1370–1390.

[46] Reinhard Pekrun, Thomas Goetz, Anne C Frenzel, Petra Barchfeld, and Raymond P Perry. 2011. Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). *Contemporary educational psychology* 36, 1 (2011), 36–48.

[47] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[48] Tianrong Rao, Xiaoxu Li, Haimin Zhang, and Min Xu. 2019. Multi-level region-based convolutional neural network for image emotion classification. *Neurocomputing* 333 (2019), 429–439.

[49] Mintra Ruensuk, Eunyong Cheon, Hwajung Hong, and Ian Oakley. 2020. How Do You Feel Online: Exploiting Smartphone Sensors to Detect Transitory Emotions during Social Media Use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–32.

[50] Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and emotion* 24, 7 (2010), 1153–1172.

[51] Jocelyn Scheirer, Raul Fernandez, and Rosalind W Picard. 1999. Expression glasses: a wearable device for facial expression recognition. In *CHI'99 Extended Abstracts on Human Factors in Computing Systems*. 262–263.

[52] Stefan Selke. 2016. *Lifelogging: Digital self-tracking and Lifelogging-between disruptive technology and cultural transformation.* Springer.

[53] Dongyu She, Jufeng Yang, Ming-Ming Cheng, Yu-Kun Lai, Paul L Rosin, and Liang Wang. 2019. Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection. *IEEE Transactions on Multimedia* 22, 5 (2019), 1358–1371.

[54] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).

[55] Paweł Tarnowski, Marcin Kołodziej, Andrzej Majkowski, and Remigiusz Jan Rak. 2020. Eye-Tracking Analysis for Emotion Recognition. *Computational Intelligence and Neuroscience* 2020 (2020).

[56] Helma Torkamaan and Jürgen Ziegler. 2020. Mobile Mood Tracking: An Investigation of Concise and Adaptive Measurement Instruments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–30.

[57] Michele M Tugade and Barbara L Fredrickson. 2004. Resilient individuals use positive emotions to bounce back from negative emotional experiences. *Journal of personality and social psychology* 86, 2 (2004), 320.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS*. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[59] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of CVPR*. 7794–7803.

[60] Yujiang Wang, Mingzhi Dong, Jie Shen, Yang Wu, Shiyang Cheng, and Maja Pantic. 2020. Dynamic Face Video Segmentation via Reinforcement Learning. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition*. 6959–6969.

[61] Hao Wu, Jinghao Feng, Xuejin Tian, Edward Sun, Yunxin Liu, Bo Dong, Fengyuan Xu, and Sheng Zhong. 2020. EMO: Real-time emotion recognition from single-eye images for resource-constrained eyewear devices. In *Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services*. 448–461.

[62] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[63] Jufeng Yang, Dongyu She, Yu-Kun Lai, Paul L Rosin, and Ming-Hsuan Yang. 2018. Weakly supervised coupled networks for visual sentiment analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7584–7592.

[64] Jufeng Yang, Dongyu She, Yu-Kun Lai, and Ming-Hsuan Yang. 2018. Retrieving and classifying affective images via deep metric learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[65] M. Zhang and Z. Zhou. 2014. A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge & Data Engineering* 26, 8 (2014), 1819–1837.

[66] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5579–5588.

[67] Mingmin Zhao, Fadel Adib, and Dina Katabi. 2016. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 95–108.

[68] Yingying Zhao, Mingzhi Dong, Yujiang Wang, Da Feng, Qin Lv, Robert P. Dick, Dongsheng Li, Tun Lu, Ning Gu, and Li Shang. 2021. A Reinforcement-Learning-based Energy-Efficient Framework for Multi-Task Video Analytics Pipeline. *IEEE Transactions on Multimedia* (2021), 1–1. https://doi.org/10.1109/TMM.2021.3076612

[69] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13041–13049.

[70] Xinge Zhu, Liang Li, Weigang Zhang, Tianrong Rao, Min Xu, Qingming Huang, and Dong Xu. 2017. Dependency exploitation: A unified CNN-RNN approach for visual emotion recognition. In *proceedings of the 26th international conf. on artificial intelligence*. 3595–3601.