EDITOR: Erik Cambria, cambria@ntu.edu.sg

DEPARTMENT: AFFECTIVE COMPUTING
AND SENTIMENT ANALYSIS

Joint Weakly-Supervised Image Emotion Analysis based on Inter-class Discrimination and Intra-class Correlation

Xinyue Zhang, *East China Normal University, Shanghai, China*

Zhaoxia Wang, *Singapore Management University, Singapore*

Guitao Cao, *East China Normal University, Shanghai, China*

Seng-Beng Ho, *Agency for Science, Technology and Research, Singapore*

Abstract—Regional information based image emotion analysis has recently garnered significant attention. However, existing methods often focus on identifying region proposals through layered steps or merely rely on visual saliency. These approaches may lead to an underestimation of emotional categories and a lack of comprehensive inter-class discrimination perception and emotional intra-class contextual mining. To address these limitations, we propose a novel approach named InterIntraIEA, which combines inter-class discrimination and intra-class correlation joint learning capabilities for image emotion analysis. The proposed method not only employs category-specific dictionary learning for class adaptation, but also models intra-class contextual relationships and perceives correlations at the channel level. This refinement process improves inter-class descriptive ability and enhances emotional categories, resulting in the production of pseudo maps that provide more precise emotional region information. These pseudo maps, in conjunction with top-level features extracted from a multi-scale extractor, are then input into a weakly-supervised fusion module to predict emotional sentiment categories. Extensive experiments conducted on four image sentiment benchmark datasets validate the superiority of our proposed method, InterIntraIEA, over state-of-the-art methods.

With the explosive growth of social media leading to a substantial increase in online image sharing, emotion analysis has garnered significant attention [1], [2]. As a crucial component of emotion analysis, image emotion analysis (IEA) aims to analyze image content to facilitate the understanding of public opinions, emotions, and cultural trends, making it an increasingly important field of study. The practical applications of IEA are extensive [3].

For example, IEA can personalize and enhance user experiences by recommending content aligned with users' emotional states or preferences [4]. Moreover, in the academic sphere, IEA spans multiple disciplines, including psychology, computer science, and linguistics, promoting interdisciplinary research and applications [5]. Traditional visual tasks aim to identify and classify visible physical elements within images, such as objects, or scenes [6], [7]. In contrast, IEA seeks to capture the emotional essence conveyed by images, which is often abstract and typically communicated through subtle hints. However, current methods of classifying emotions based on regional information face the following challenges.

Digital Object Identifier 10.1109/MIS.2024.XXXXXXX

Firstly, while the emotion regions detected in most images directly convey sentiments, other categories of emotional information containing context should not be overlooked. Context provides additional insights for analyzing emotions, particularly when they are vague, and improves model robustness by diversifying the features used for emotion recognition. Secondly, although various visual saliency based approaches have been developed to highlight the relative importance of different areas, due to the inherent subjectivity and ambiguity of human emotions [8], [9], [10], the semantic features of each class are likely to be intertwined. Emotion regions identified solely based on visual saliency might not accurately represent the intended emotions.

Therefore, we propose a joint weakly-supervised learning network named InterIntraIEA, which integrates inter-class discrimination and intra-class correlation to tackle challenges in region-based IEA research. This approach enables the model to disentangle emotional categories and understand the context of emotion regions. Firstly, leveraging the advantages of data aggregation, we design an inter-class discrimination sub-module. This sub-module, utilizing a class-specific dictionary, learns scaling factors for spatial features, encoding each category to alleviate entanglement and improve recognition of emotion regions. Secondly, drawing inspiration from visual saliency's top-down approach, we develop an intra-class correlation sub-module. This sub-module establishes interactions and connections between specific features expressing emotions and their context, focusing on pivotal features for predicting emotional categories. Subsequently, in the pseudo map generation process, we integrate the outputs of the inter-class discrimination and intra-class correlation sub-modules to generate more accurate pseudo sentiment maps. Finally, these pseudo maps, combined with top-level features from the multi-scale extractor, are inputted into a weakly-supervised fusion module for predicting emotion categories.

Overall, our research contributions mainly encompass the following four aspects:

- We propose a novel approach, InterIntraIEA, which integrates inter-class discrimination and intra-class correlation joint learning capabilities for analyzing emotional sentiment in images.
- The proposed InterIntraIEA integrates a joint learning module for analyzing both inter- and intra-class emotional feature representations, distinguishing between emotion categories and capturing contextual correlations. This dual-focus approach allows for precise identification of emotion regions.
- The proposed InterIntraIEA utilizes a weakly-supervised fusion module that integrates pseudo maps from the joint learning module, and top-level features extracted from a multi-scale extractor, to predict emotional categories.
- Experimental results across four distinct datasets demonstrate the superior performance of the proposed InterIntraIEA compared to existing state-of-the-art methods. These results not only showcase the effectiveness of the proposed InterIntraIEA, but also validate its practicality and significant advancement in the field of IEA .

RELATED WORK

In the domain of IEA , pioneering researchers initially focused on constructing hand-crafted features at various levels: low, mid, and high [12], for the categorization and understanding of emotions expressed by images. As deep learning has advanced, a multitude of methodologies exploit deep neural networks to independently learn and derive features, marking a significant shift from manual feature engineering. Borrowed from object detection, the integration of region proposals represents an innovative stride. For instance, Zhang et al. [13] harnessed region detectors to unearth multi-level region proposals for nuanced recognition. Rao et al. [14] proposed a specific method for generating emotionally significant local regions tailored for sentiment recognition. The essence of these approaches lies in pinpointing sentiment-rich regions within images, thereby boosting the efficiency and accuracy of IEA .

Despite their advancements, those methods share a common hurdle: the derivation of region proposals entails intricate computational efforts, presenting an opportunity for weakly supervised based approaches to make a significant impact. Prevalent IEA frameworks generally rely on weakly supervised strategies [15]. In light of recent studies, some scholars introduced the human cognitive mechanism [16] into IEA [17]. However, challenges persist in separating semantic features across categories, often causing class-specific trait overlap. InterIntraIEA adopts class-specific dictionary learning beyond saliency information, improving the network's ability to differentiate emotional categories.

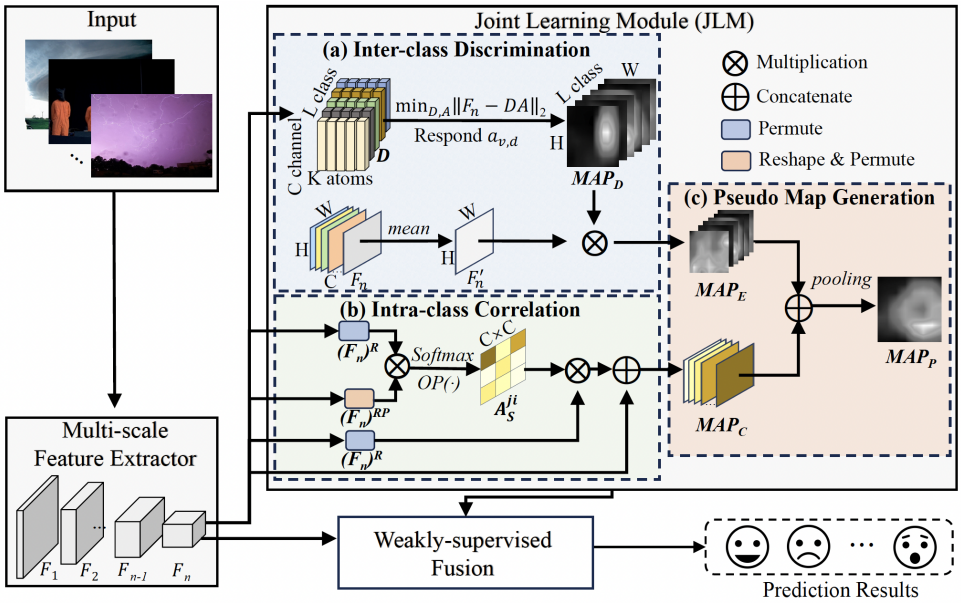

FIGURE 1. Illustration of the pipeline of InterIntraIEA: Images undergo initial processing using a backbone network (Res2Net-101 [11]) as a multi-scale feature extractor. Subsequently, the output is directed to a Joint Learning Module (JLM) consisting of three sub-modules: (a) Inter-Class Discrimination for emotion category discrimination, (b) Intra-Class Correlation for context correlation awareness, and (c) Pseudo Map Generation, utilized to generate pseudo emotional maps. Finally, these pseudo maps, combined with top-level features from the multi-scale extractor, are inputted into a weakly-supervised fusion module for predicting emotion categories.

THE METHODOLOGY

Overview

As illustrated in Figure 1, InterIntraIEA contains (i) a multi-scale feature extractor that fully utilizes the multi-scale features of images, employing the Res2Net-101 backbone network [11] to extract features at different levels, which comprises n convolutional blocks $\{F_1, F_2, ..., F_n\}$. (ii) a novel joint learning module, which encompasses three distinct sub-modules: the inter-class discrimination sub-module, aims to differentiate among various emotional categories and extracts emotion category discrimination; the intra-class correlation sub-module focuses on identifying contextual links within the emotional channels, building context correlation awareness;and conclusively, the pseudo map generation sub-module capitalizes on the synergistic qualities of inter-class and intra-class dynamics, orchestrating the precise identification and localization of pivotal regions that trigger the predominant emotion. (iii) a weakly-supervised fusion module that integrates pseudo maps with the top-level feature maps to weakly supervise the final emotion classification, thereby enhancing the overall performance of the IEA task.

Joint Learning Module

As illustrated in Figure 1, the joint learning module (JLM) consists of three components: the inter-class discrimination sub-module (depicted in Figure 1(a)), the intra-class correlation sub-module (shown in Figure 1(b)), and the pseudo map generation sub-module (presented in Figure 1(c)). Drawing on the theory of visual attention, we introduce the intra-class correlation sub-module, which enhances the representation of emotion-related semantic features and constructs an emotion category-aware attention map, aiming to grasp the subtle intra-class relationships. The inter-class discrimination sub-module, by establishing an emotion category dictionary, encodes specific categories as a linear combination of a set of basis vectors based on the emotional dictionary, thereby highlighting the inter-class differences. After learning inter-class discrimination and intra-class correlation, the pseudo map generation sub-module employs a customized pooling strategy to generate pseudo maps that precisely reveal emotion regions within images.

Inter-class Discrimination

We construct an inter-class discrimination mechanism by encoding explicit class semantic information into

class attention maps for each atomic group. We build a learned L class dictionary $D = \{d_1, ...d_i, ..., d_{L\times M}\}$, $d_i \in R^C$. Here M represents the number of atoms for each category. We use the dictionary D and sparse coefficients $A = \{a_1, ...a_i\}$ to represent the learned feature F_n, transitioning it from color space to sparse space, which can be formulated as $min_{D,A} \|F_n - DA\|_2$.

To solve the optimization problem of the $min_{D,A} \|F_n - DA\|_2$, we perform similarity computation between a pixel vector v_i from the feature in F_n and the j-th class atom vector d_j in D in the original space using the inner product kernel function $k(v_i, d_j) = ((v_i)^T d_j)^2$. We then obtain the response a_{vd} of v_i on d_j through Equation (1).

$$a_{vd} = \frac{k(v_i, d_j)}{\sum_{l=1}^{L\times M} k(v_i, d_j)} \tag{1}$$

Here, we construct the kernel function $k(v_i, d_j) = \exp(-d_j^T v_i)$, where $f(t) = e^t$, $-\infty < t < \infty$. As $f^{(n)}(t) = e^t > 0$, $k(v_i, d_j)$ is a kernel function. Therefore, the response matrices can be represented as $A \in R^{L\times M\times H\times W}$. To obtain the class-specific guidance maps, we perform an average pooling operation on the second dimension of A, resulting in $MAP_D \in R^{L\times H\times W}$. Considering reducing the computational complexity, we utilize a channel-wise average pooling operation to reduce the dimensionality of $F_n \in R^{C\times H\times W}$ to $F_n' \in R^{1\times H\times W}$. Finally, we multiply MAP_D and F_n' to obtain the MAP_E^l as Equation (2):

$$MAP_E^l = MAP_D \otimes F_n' \tag{2}$$

where $\otimes$ represents the multiplication operation. Then we concatenate $l \in L$ categories of MAP_E^l to obtain the output MAP_E of this sub-module.

Intra-class Correlation

We utilize intra-class correlation sub-module to model channel correlations, adaptively aggregating contextual information, thereby enhancing the representation of emotion-related features. We leverage the features F_n generated by the last convolutional block in the multi-scale feature extractor as the input. In the context awareness attention, we perform reshape and permute operations on F_n, resulting in $(F_n)^R \in R^{C\times N}(N = W \times H)$ and $(F_n)^P$, respectively. To capture the channel dependencies between any two positions within F_n, we first calculate the matrix multiplication result A_m of the enhanced matrices $(F_n)^R$ and $(F_n)^{RP}$: $A_m = (F_n)^R \otimes (F_n)^{RP}$, where $\otimes$ represents the matric multiplication. Then we apply the $OP(\cdot)$ operation to suppress features that are less prominent or have lower values, and make emotion-related features more easily interpretable by subsequent layers of the network: $OP(A_m) = f_M(A_m, -1) - A_m$, where f_M represents the function that takes the maximum value between A_m and -1. We adopt the Equation (3) below to obtain a $C \times C$ adjacency matrix, which reallocates weights to emotion-related features, helping the network to focus on more significant emotional features:

$$A_S^{ji} = \frac{exp(OP(A_m^i \cdot A_m^j))}{\sum_{i=1}^{C} exp(OP(A_m^i \cdot A_m^j))} \tag{3}$$

where A_m^i represents the i-th channel, while A_S^{ji} represents the influence of the i-th channel on the j-th channel in the attention map. Finally, the output of the context-awareness attention is obtained by the following formula:

$$MAP_C = \theta F_n + (\sum_{i=1}^{C} A_S^{ji} \otimes (F_n)^R) \tag{4}$$

where θ represents a learnable scale factor that is initialized to zero.

Pseudo Map Generation

After obtaining MAP_E from inter-class discrimination sub-module and MAP_C from the intra-class correlation sub-module, we calculate the weight w_l for image-level pseudo emotion label as follows:

$$w_l = \frac{1}{n} \sum_{i=1}^{n} g_{GAP} \{MAP_{IT}(i, l)\} \tag{5}$$

where we utilize n emotional class-related detectors to generate w_l. g_{GAP} represents the global average pooling function. $\{MAP_{IT}(i, l)\}$ refers to an interaction between feature maps and emotion categories, that the corresponding i-th feature map of the l-th emotional label. Then we leverage the pooling strategy $g_{pooling}$ (shown in Equation (6)) to obtain the the emotional region map MAP_P, which serves as the pseudo map for the entire weakly-supervised framework.

$$g_{pooling} = \sum_{l=1}^{L} (\frac{1}{m} \sum_{i=1}^{m} MAP_{IT}(i, l)) w_l \tag{6}$$

Weakly-supervised Fusion Module

InterIntraIEA first highlights emotion-related regions through the JLM, thereby enhancing the classification effect and generating pseudo maps to guide the prediction for multi-class emotions [18]. Therefore, we derive the final prediction Pre with MAP_P from JLM module and F_n from the last convolutional block of multi-scale feature extractor:

$$Pre = f_{st}(g_{GAP}(concat(MAP_P, F_n))) \quad (7)$$

where f_{st} denotes the *Softmax* function, and concat(MAP_P,F_n) represents the concatenate operation for MAP_P and F_n.

EXPERIMENTAL EVALUATION

Datasets

We've leveraged four datasets to validate InterIntraIEA's performance across different emotional contexts. We engage with both large-scale public datasets and more specific affective collections: one large dataset, Flickr and Instagram (FI-8) [19], and three additional widely recognized datasets: EmotionROI (6 classes) [15], IAPS-Subset [20], and Twitter II [15].

Implementation Details

The entire implementation was employed on the PyTorch 1.2.0 framework. The input images were resized to 448 $\times$ 448 pixels for uniformity, and then through a combination of random crop and horizontal flips to enhance the variety of the training set. During the training phase, given the role of InterIntraIEA in tackling tasks involving multiple emotion classifications, the Cross Entropy Loss was utilized for both the pseudo map generation process and the weakly-supervised fusion module. We selected Stochastic Gradient Descent for optimization. The values of momentum and weight decay rates were set to 0.9 and 0.0005, respectively. The batch size was set to 12, and the learning rate was initialized to 0.0001. During the testing phase, the model is conducted three times, and the average of these results is reported as InterIntraIEA's overall performance. The experiments were performed on an Nvidia Tesla P100-PCIE with 16GB on-board memory.

Comparison with Different Methods

We evaluate the performance of InterIntraIEA on the extensive FI-8 dataset compared to various frameworks in Table 1, and on smaller-scale datasets as shown in Table 2. Against baseline approaches, InterIntraIEA shows enhancements in evaluation metric accuracy. For the large-scale dataset (shown in Table 1), InterIntraIEA surpasses the state-of-the-art methods by Yang et al. [21] and DCNet [17], improving performance by 1.71% and 1.19%, respectively. Large datasets often contain noise, however, results indicate that InterIntraIEA can effectively handle the inevitable noise within large-scale datasets. For small-scale datasets (shown in Table 2), Yamamoto et al. [19] combine visual and semantic features of emotion regions to train a support vector machine emotion classifier. Yang et al. [21] employs a feature fusion, while DCNet [17] integrates high-level and low-level features to identify emotionally significant areas to guide emotion classification. InterIntraIEA emphasizes leveraging human visual attention for solving challenges in IEA , and focuses on identifying semantic features of emotion categories from a class-specific encoding perspective, enhancing accuracy of IEA by integrating this with visual attention metrics, which shows improvements over the state-of-the-art method DCNet: a 0.37% on EmotionROI, 0.12% on IAPS-Subset, and 0.56% on Twitter II.

TABLE 1. Comparison with different methods on FI-8 dataset.

Methods	Publication Year	FI-8
Zhao's [22]	2014	46.13
Sentibank [12]	2013	49.23
DeepSentibank [23]	2014	51.54
ImageNet-AlexNet	2017	38.26
ImageNet-VGG16	2014	41.22
ImageNet-ResNet101	2016	50.01
Yang's [24]	2017	66.79
WILDCAT [25]	2017	67.03
CAM [26]	2016	68.54
WSCNet [15]	2019	70.07
Yamamoto's [19]	2021	70.46
Yang's [21]	2023	71.13
DCNet [17]	2023	71.65
InterIntraIEA		**72.84**

Additionally, we present the confusion matrices for FI-8 in Figure 2, and for two smaller datasets in Figure 3. InterIntraIEA performs well on both multi-class and binary datasets, despite some confusion between categories. For example, in Figure 2, Disgust is easily confused with other classes, which we attribute to the high feature overlap in the large FI-8 dataset, making distinction more challenging and potentially leading to confusion. In Figure 3, such issues are mitigated in the two smaller datasets categorized into positive and negative emotions. The clear dichotomy between these two categories simplifies distinction, reducing the complexity of model recognition and classification. Overall, the correct identification rate surpasses the confusion rate with accurate categories.

Ablation Studies for JLM

Since WSCNet [15] is a classic and effective method in this field, and DCNet [17] represents state-of-the-art accuracy, we regard them as baseline benchmarks. We conducted a comprehensive evaluation of three methods (WSCNet, DCNet, and InterIntraIEA) across two distinct datasets (FI-8 and EmotionROI), which

TABLE 2. Performance comparison on three small-scale datasets using accuracy as the metric.

Methods	EmotionROI	Methods	IAPS-Subset	Methods	Twitter II
Zhao's [22]	34.84	SentiBank [12]	81.79	DeepSentiBank [23]	70.23
DeepSentiBank [23]	42.53	DeepSentiBank [23]	85.63	VGGNet	71.79
Yang's [24]	52.40	PCNN [27]	88.84	WILDCAT [25]	78.81
WILDCAT [25]	55.05	VGGNet	88.51	CAM [26]	79.13
CAM [26]	55.72	Yang's [28]	92.39	Sun's [29]	80.91
WSCNet [15]	58.25	Zhang's [20]	95.83	WSCNet [15]	81.35
DCNet [17]	59.60	DCNet [17]	95.90	DCNet [17]	82.50
InterIntraIEA	**59.97**	InterIntraIEA	**96.02**	InterIntraIEA	**83.06**

FIGURE 2. Confusion matrix on large-scale FI-8 dataset.

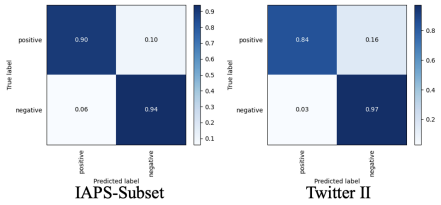

FIGURE 3. Confusion matrix on two binary datasets.

components, the proposed model, upon their integration, exhibited more significant performance improvements on both large and small datasets, particularly on EmotionROI, highlighting InterIntraIEA's potential in understanding and analyzing more complex emotional scenarios. Furthermore, our analysis revealed that InterIntraIEA, incorporating intra-class correlation and inter-class discrimination in all tested configurations, not only improved overall accuracy but also achieved a more balanced recognition rate across different emotion categories.

TABLE 3. Impact of the proposed Joint Learning Module.

	Intra-class	Inter-class	FI-8	EmotionROI
WSCNet	✗	✗	70.07	58.25
	✓	✗	70.51	58.80
	✗	✓	70.59	58.84
	✓	✓	**71.06**	**59.15**
DCNet	✗	✗	71.65	59.60
	✓	✗	72.23	59.73
	✗	✓	72.30	59.75
	✓	✓	**72.55**	**59.92**
Ours	✓	✗	71.87	59.01
	✗	✓	72.25	59.42
	✓	✓	**72.84**	**59.97**

is demonstrated in Table 3. We observed that the inclusion of intra-class correlation and inter-class discrimination, both individually and combined, not only universally enhances the performance of models in emotion recognition tasks but also reveals a significant synergistic effect on performance improvement when these components are integrated. Notably, the proposed method, with both components integrated, achieved the highest accuracy rates on both datasets (72.84% on FI-8 and 59.97% on EmotionROI), underscoring the efficacy of combining intra-class correlation and inter-class discrimination to enhance emotion recognition precision. Moreover, although DCNet showed higher accuracy before the integration of these

Comparison of Different Pseudo Maps

In Figure 4, we visualize the pseudo maps generated by different methods, which highlight crucial areas that expose underlying emotions. Specifically, we compare the saliency maps generated based on visual saliency theory [30] (column 2) with different pseudo maps produced by CAM [26] (column 3), DCNet [17] (column 4), and InterIntraIEA (column 5). In simple scenes, such as the first row where a lady is covering her face while crying, saliency maps roughly outline the lady's figure, while the CAM method identifies the hands as the main region for classification. In contrast, pseudo maps from InterIntraIEA, after intra-class correlation and inter-class discrimination processes, locate emotion regions more accurately than other methods. In

the second and third rows, saliency maps fail to pinpoint salient regions. While both the CAM method and DCNet identify emotion regions, they remain somewhat vague. InterIntraIEA, however, can determine the final emotion regions more clearly in complex scenes.

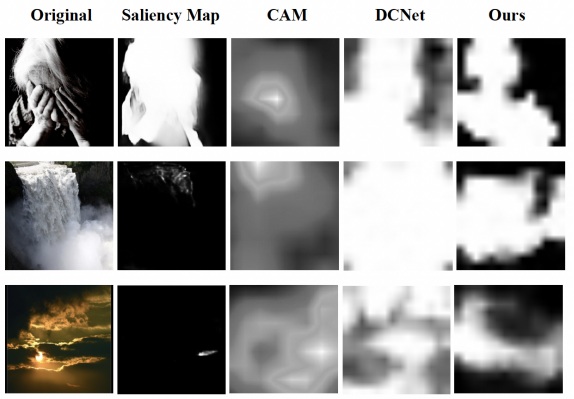

FIGURE 4. Visualization for pseudo maps generated by different methods.

CONCLUSION

In conclusion, this paper delves into the branch of IEA focusing on regional information-based approaches, addressing the limitations inherent in existing methods reliant on region proposals or visual saliency. We have developed a weakly-supervised framework built around a joint learning module that effectively employs category-specific dictionary learning to improve class adaptation and models the intra-class contextual relationships of emotional categories. This approach not only strengthens the discriminative capability between classes but also refines emotional categories, leading to a more precise identification of emotion regions through the pseudo map generation process. For future research directions, we aim to delve into the integration of multimodal data, incorporating textual information alongside visual cues.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China under Grant 61871186, 61771322, and we sincerely appreciate the support from the China Scholarship Council.

REFERENCES

1. E. Cambria et al., “Seven pillars for the future of artificial intelligence,” *IEEE Intelligent Systems*, vol. 38, no. 6, 2023, pp. 62–69.
2. L. Stappen et al., “Sentiment analysis and topic recognition in video transcriptions,” *IEEE Intelligent Systems*, vol. 36, no. 2, 2021, pp. 88–95.
3. J. Cui et al., “Survey on sentiment analysis: evolution of research methods and topics,” *Artificial Intelligence Review*, vol. 56, 2023, pp. 8469–8510.
4. X. Liu and K. Lee, “Optimized facial emotion recognition technique for assessing user experience,” *2018 IEEE Games, Entertainment, Media Conference (GEM)*, 2018, pp. 1–9.
5. E. Cambria et al., “Sentiment Analysis is a big suitcase,” *IEEE Intelligent Systems*, vol. 32, no. 6, 2017, pp. 74–80.
6. C. Tang et al., “Visual recognition by request,” *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15265–15274.
7. B. Mahaur, K. Mishra, and N. Singh, “Improved Residual Network based on norm-preservation for visual recognition,” *Neural Networks*, vol. 157, 2023, pp. 305–322.
8. Z. Wang et al., “MiMuSA—mimicking human language understanding for fine-grained multi-class sentiment analysis,” *Neural Computing and Applications*, vol. 35, 2023, pp. 15907–15921.
9. E. Cambria, D. Olsher, and K. Kwok, “Sentic activation: A two-level affective common sense reasoning framework,” *Proceedings of the AAAI Conference on Artificial Intelligence, vol. 26*, 2012, pp. 186–192.
10. F. Z. Xing, F. Pallucchini, and E. Cambria, “Cognitive-inspired domain adaptation of sentiment lexicons,” *Information Processing & Management*, vol. 56, no. 3, 2019, pp. 554–564.
11. S.-H. Gao et al., “Res2Net: A New Multi-scale Backbone Architecture,” *IEEE TPAMI*, 2021, doi:10.1109/TPAMI.2019.2938758.
12. D. Borth et al., “Large-scale visual sentiment ontology and detectors using adjective noun pairs,” *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 223–232.
13. J. Zhang et al., “Image sentiment classification via multi-level sentiment region correlation analysis,” *Neurocomputing*, vol. 469, 2022, pp. 221–233.
14. T. Rao et al., “Multi-level region-based convolutional neural network for image emotion classification,” *Neurocomputing*, vol. 333, 2019, pp. 429–439.
15. D. She et al., “Wscnet: Weakly supervised coupled networks for visual sentiment classification and de-

tection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, 2019, pp. 1358–1371.
16. E. Cambria et al., "SenticNet 8: Fusing Emotion AI and Commonsense AI for Interpretable, Trustworthy, and Explainable Affective Computing," *International Conference on Human-Computer Interaction (HCII)*, 2024.
17. X. Zhang et al., "DCNet: Weakly Supervised Saliency Guided Dual Coding Network for Visual Sentiment Recognition," *26th European Conference on Artificial Intelligence*, 2023, pp. 3050 – 3057.
18. Z. Wang, S.-B. Ho, and E. Cambria, "A review of emotion sensing: categorization models and algorithms," *Multimedia Tools and Applications*, vol. 79, 2020, pp. 35553–35582.
19. T. Yamamoto, S. Takeuchi, and A. Nakazawa, "Image emotion recognition using visual and semantic features reflecting emotional and similar objects," *IEICE TRANSACTIONS on Information and Systems*, vol. 104, no. 10, 2021, pp. 1691–1701.
20. H. Zhang and M. Xu, "Weakly supervised emotion intensity prediction for recognition of emotions in images," *IEEE Transactions on Multimedia*, vol. 23, 2020, pp. 2033–2044.
21. H. Yang et al., "Exploiting emotional concepts for image emotion recognition," *The Visual Computer*, vol. 39, no. 5, 2023, pp. 2177–2190.
22. S. Zhao et al., "Exploring principles-of-art features for image emotion recognition," *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 47–56.
23. T. Chen et al., "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.
24. J. Yang, D. She, and M. Sun, "Joint Image Emotion Classification and Distribution Learning via Deep Convolutional Neural Network." *IJCAI*, 2017, pp. 3266–3272.
25. T. Durand et al., "Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 642–651.
26. B. Zhou et al., "Learning deep features for discriminative localization," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.
27. Q. You et al., "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks," *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, AAAI Press, 2015, ISBN 0262511290, pp. 381–388.
28. J. Yang et al., "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Transactions on Multimedia*, vol. 20, no. 9, 2018, pp. 2513–2525.
29. M. Sun et al., "Discovering affective regions in deep convolutional neural networks for visual sentiment prediction," *2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.
30. S. Chen et al., "Reverse attention-based residual network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, 2020, pp. 3763–3776.

Xinyue Zhang is currently pursuing a Ph.D. degree at East China Normal University under the supervision of Professor Guitao Cao. She is also a visiting Ph.D. student at Singapore Management University under the supervision of Associate Professor Zhaoxia Wang. Her research interests include visual sentiment recognition and AI in education. Contact her at xyzhang@stu.ecnu.edu.cn.

Zhaoxia Wang is an Associate Professor of Computer Science (Practice) in the School of Computing and Information Systems of Singapore Management University. Her research interests include natural language processing, machine learning, sentiment analysis, causal reasoning, and intelligent robots. Contact her at zxwang@smu.edu.sg.

Guitao Cao is currently a Professor of Software Engineering Institute, East China Normal University, Shanghai. She has published decades of peer reviewed papers in top venues including IEEE Transactions on Cybernetics, IEEE Transactions on Multimedia, and IEEE Transactions on Biomedical Engineering. Her research interests include machine learning and its application, image understanding and analysis, and edge computing. Contact her at gtcao@sei.ecnu.edu.cn.

Seng-Beng Ho is a Principal Scientist at the Institute of High Performance Computing, A*STAR, Singapore. He has published many AI-related papers in international journals and conferences in which he presents principled and fundamental theoretical frameworks that are critical for building truly general AI systems. Contact him at hosb@ihpc.a-star.edu.sg.