10-2024

Joint weakly supervised image emotion analysis based on interclass discrimination and intraclass correlation

Xinyue ZHANG

Zhaoxia WANG
Singapore Management University, zxwang@smu.edu.sg

Guitao CAO

Seng-Beng HO

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

DEPARTMENT: AFFECTIVE COMPUTING AND SENTIMENT ANALYSIS

Joint Weakly Supervised Image Emotion Analysis Based on Interclass Discrimination and Intraclass Correlation

Xinyue Zhang, *East China Normal University, Shanghai, 200062, China*
Zhaoxia Wang, *Singapore Management University, 178902, Singapore*
Guitao Cao, *East China Normal University, Shanghai, 200062, China*
Seng-Beng Ho, *AI Institute Global, 139955, Singapore*

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 interclass discrimination perception and emotional intraclass contextual mining. To address these limitations, we propose a novel approach named InterIntraIEA, which combines interclass discrimination and intraclass correlation joint learning capabilities for image emotion analysis. The proposed method not only employs category-specific dictionary learning for class adaptation, but also models intraclass contextual relationships and perceives correlations at the channel level. This refinement process improves interclass descriptive ability and enhances emotional categories, resulting in the production of pseudomaps that provide more precise emotional region information. These pseudomaps, in conjunction with top-level features extracted from a multiscale extractor, are then input into a weakly supervised fusion module to predict emotional sentiment categories.

With the explosive growth of social media leading to a substantial increase in online image sharing, emotion analysis has garnered significant attention.[1]

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.

1541-1672 © 2024 IEEE
Digital Object Identifier 10.1109/MIS.2024.3441408
Date of current version 2 October 2024.

For example, IEA can personalize and enhance user experiences by recommending content aligned with users' emotional states or preferences.[2] Moreover, in the academic sphere, IEA spans multiple disciplines, including psychology, computer science, and linguistics, promoting interdisciplinary research and applications.[3] Traditional visual tasks aim to identify and classify visible physical elements within images, such as objects, or scenes. 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.

First, 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. Second, 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,[4] 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 interclass discrimination and intraclass 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. First, leveraging the advantages of data aggregation, we design an interclass discrimination submodule. This submodule, utilizing a class-specific dictionary, learns scaling factors for spatial features, encoding each category to alleviate entanglement and improve recognition of emotion regions. Second, drawing inspiration from visual saliency's top-down approach, we develop an intraclass correlation submodule. This submodule establishes interactions and connections between specific features expressing emotions and their context, focusing on pivotal features for predicting emotional categories. Subsequently, in the pseudomap generation process, we integrate the outputs of the interclass discrimination and intraclass correlation submodules to generate more accurate pseudosentiment maps. Finally, these pseudomaps, combined with top-level features from the multiscale 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 interclass discrimination and intraclass correlation joint learning capabilities for analyzing emotional sentiment in images.
- The proposed InterIntraIEA integrates a joint learning module for analyzing both inter- and intraclass 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 pseudomaps from the joint learning module, and top-level features extracted from a multiscale 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,[6] 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.[7] harnessed region detectors to unearth multilevel region proposals for nuanced 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.[8] In light of recent studies, some scholars introduced the human cognitive mechanism into IEA.[9] 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.

THE METHODOLOGY

Overview

As illustrated in Figure 1, InterIntraIEA contains 1) a multiscale feature extractor that fully utilizes the multiscale features of images, employing the Res2Net-101 backbone network[5] to extract features at different

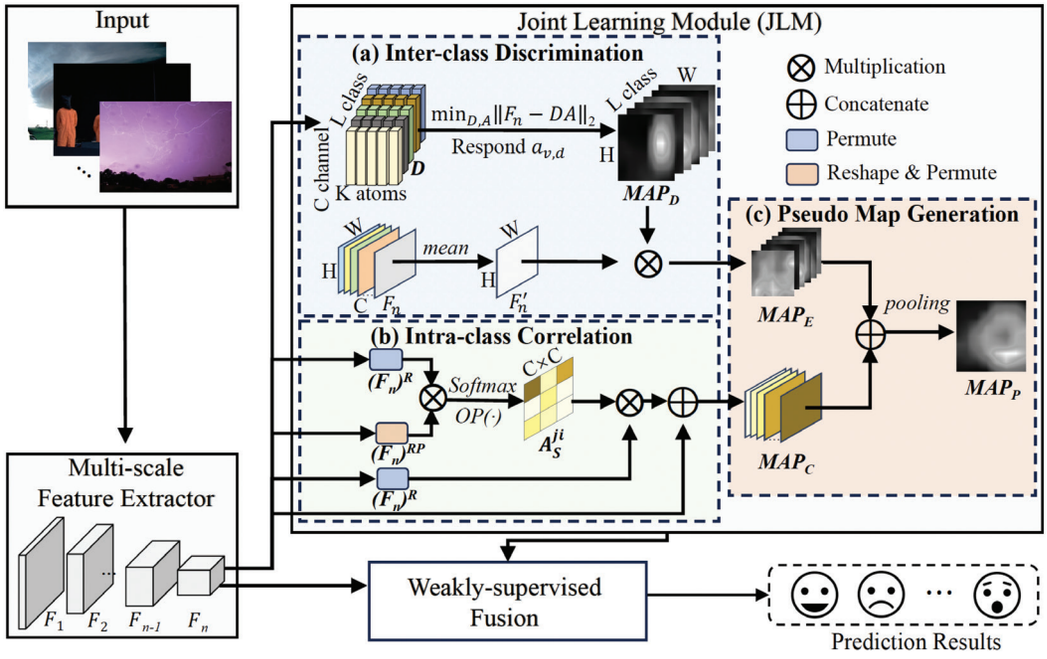

FIGURE 1. Illustration of the pipeline of InterIntraIEA: Images undergo initial processing using a backbone network (Res2Net-101[5]) as a multiscale feature extractor. Subsequently, the output is directed to a JLM consisting of three submodules: (a) interclass discrimination for emotion category discrimination, (b) intraclass correlation for context correlation awareness, and (c) pseudomap generation, utilized to generate pseudoemotional maps. Finally, these pseudomaps, combined with top-level features from the multiscale extractor, are inputted into a weakly supervised fusion module for predicting emotion categories.

levels, which comprises n convolutional blocks $\{F_1, F_2, ..., F_n\}$; 2) a novel joint learning module, which encompasses three distinct submodules: the interclass discrimination submodule, aims to differentiate among various emotional categories and extracts emotion category discrimination; the intraclass correlation submodule focuses on identifying contextual links within the emotional channels, building context correlation awareness; and conclusively, the pseudomap generation submodule capitalizes on the synergistic qualities of interclass and intraclass dynamics, orchestrating the precise identification and localization of pivotal regions that trigger the predominant emotion; 3) a weakly supervised fusion module that integrates pseudomaps 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 interclass discrimination submodule [depicted in Figure 1(a)], the intraclass correlation submodule [shown in Figure 1(b)], and the pseudomap generation submodule [presented in Figure 1(c)]. Drawing on the theory of visual attention, we introduce the intraclass correlation submodule, which enhances the representation of emotion-related semantic features and constructs an emotion category-aware attention map, aiming to grasp the subtle intraclass relationships. The interclass discrimination submodule, 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 interclass differences. After learning interclass discrimination and intraclass correlation, the pseudomap generation submodule employs a customized pooling strategy to generate pseudomaps that precisely reveal emotion regions within images.

Interclass Discrimination

We construct an interclass 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 jth 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 as follows:

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

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 $\mathrm{MAP}_D \in R^{L \times H \times W}$. 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 follows:

$$\mathrm{MAP}_E^l = \mathrm{MAP}_D \otimes F_n' \tag{2}$$

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 submodule.

Intraclass Correlation

We utilize intraclass correlation submodule 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 multiscale 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 (3) 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 as follows:

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

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

$$\mathrm{MAP}_C = \theta F_n + \left(\sum_{i=1}^{C} A_S^{ji} \otimes (F_n)^R \right) \tag{4}$$

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

Pseudo Map Generation

After obtaining MAP_E from interclass discrimination submodule and MAP_C from the intraclass correlation submodule, we calculate the weight w_l for image-level pseudoemotion label as follows:

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

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

$$g_{\mathrm{Pooling}} = \sum_{l=1}^{L} \left(\frac{1}{m} \sum_{i=1}^{m} \mathrm{MAP}_{IT}(i, l) \right) w_l. \tag{6}$$

Weakly Supervised Fusion Module

InterIntraIEA first highlights emotion-related regions through the JLM, thereby enhancing the classification effect and generating pseudomaps to guide the prediction for multiclass emotions. Therefore, we derive the final prediction *Pre* with MAP_P from JLM module and F_n from the last convolutional block of multiscale feature extractor:

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

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 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),[10] and three additional widely recognized datasets: EmotionROI (6 classes),[8] IAPS-Subset,[9] and Twitter II.[8]

Implementation Details

The entire implementation was employed on the PyTorch 1.2.0 framework. The input images were resized to 448×448 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 16-GB onboard memory.

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

Methods	Publication Year	FI-8
ImageNet-VGG16	2014	41.22
ImageNet-ResNet101	2016	50.01
CAM[11]	2016	68.54
ImageNet-AlexNet	2017	38.26
WSCNet[8]	2019	70.07
Yamamoto's[10]	2021	70.46
Yang's[6]	2023	71.13
DCNet[9]	2023	71.65
InterIntraIEA		**72.84**

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.[6] and DCNet,[9] 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.[10] combine visual and semantic features of emotion regions to train a support vector machine emotion classifier. Yang et al.[6] employs a feature fusion, while DCNet[9] 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.

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 multiclass 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

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

Methods	EmotionROI	Methods	IAPS-Subset	Methods	Twitter II
Yang's[12]	52.40	VGGNet	88.51	VGGNet	71.79
CAM[11]	55.72	Yang's[13]	92.39	CAM[11]	79.13
WSCNet[8]	58.25	Zhang's[14]	95.83	WSCNet[8]	81.35
DCNet[9]	59.60	DCNet[9]	95.90	DCNet[9]	82.50
InterIntraIEA	**59.97**	InterIntraIEA	**96.02**	InterIntraIEA	**83.06**

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[8] is a classic and effective method in this field, and DCNet[9] 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 is demonstrated in Table 3. We observed that the inclusion of intraclass correlation and interclass 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 intraclass correlation and interclass discrimination to enhance emotion recognition precision. Moreover, although DCNet showed higher accuracy before the integration of these 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

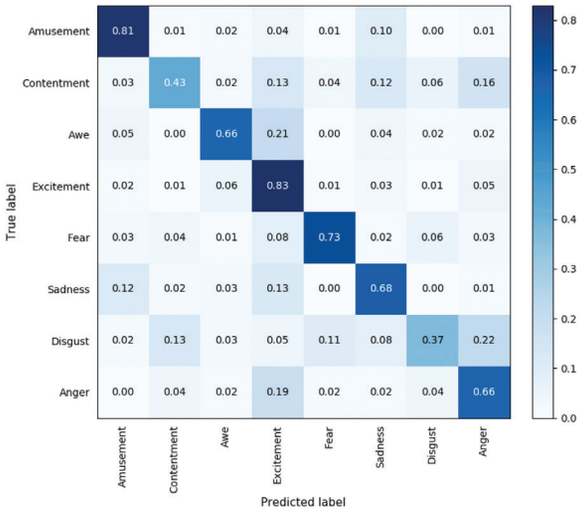

FIGURE 2. Confusion matrix on large-scale FI-8 dataset.

TABLE 3. Impact of the proposed JLM.

	Intraclass	Interclass	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**

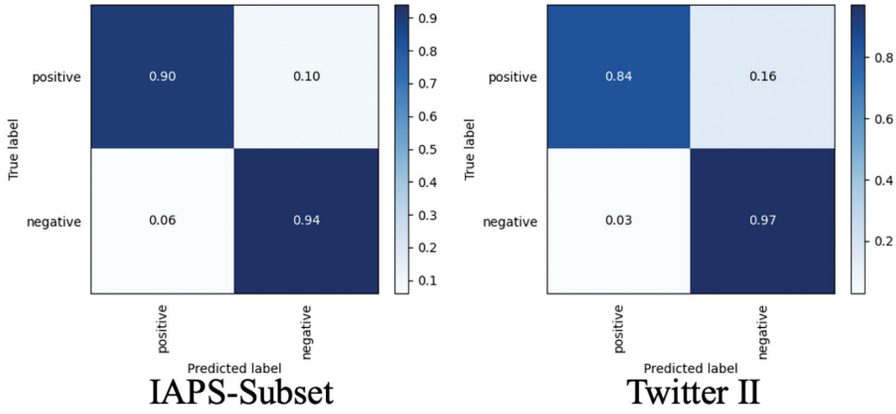

FIGURE 3. Confusion matrix on two binary datasets.

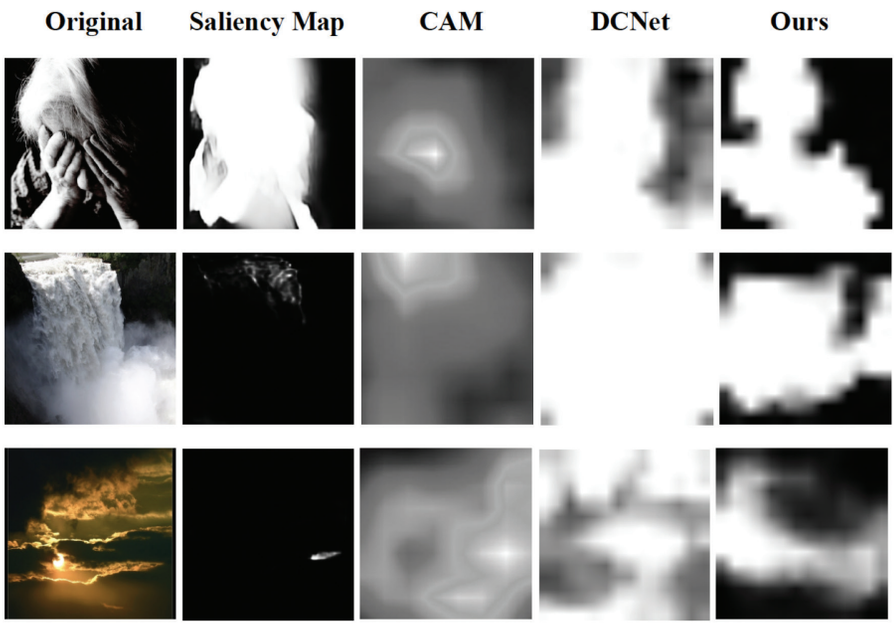

FIGURE 4. Visualization for pseudomaps generated by different methods.

intraclass correlation and interclass discrimination in all tested configurations, not only improved overall accuracy but also achieved a more balanced recognition rate across different emotion categories.

Comparison of Different Pseudo Maps

In Figure 4, we visualize the pseudomaps generated by different methods, which highlight crucial areas that expose underlying emotions. Specifically, we compare the saliency maps generated based on visual saliency theory[15] (column 2) with different pseudomaps produced by class activation mapping (CAM)[11] (column 3), DCNet[9] (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, pseudomaps from InterIntraIEA, after intraclass correlation and interclass 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.

CONCLUSION

In conclusion, this article 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 intraclass 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 pseudomap 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 Grants 61871186 and 61771322. Xinyue Zhang also acknowledges the support of the China Scholarship Council program (Project ID: 202306140115).

REFERENCES

1. E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intell. Syst.*, vol. 38, no. 6, pp. 62–69, 2023, doi: 10.1109/MIS.2023.3329745.
2. Z. Wang, S.-B. Ho, and E. Cambria, "A review of emotion sensing: Categorization models and algorithms," *Multimedia Tools Appl.*, vol. 79, nos. 47–48, pp. 35,553–35,582, 2020, doi: 10.1007/s11042-019-08328-z.

3. E. Cambria et al., "Sentiment analysis is a big suitcase," *IEEE Intell. Syst.*, vol. 32, no. 6, pp. 74–80, Nov./Dec. 2017, doi: 10.1109/MIS.2017.4531228.
4. E. Cambria et al., "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *Proc. Int. Conf. Human-Comput. Interact. (HCII)*, 2024, pp. 1–20.
5. S.-H. Gao et al., "Res2Net: A new multi-scale backbone architecture," *Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 2, pp. 652–662, Feb. 2021, doi: 10.1109/TPAMI.2019.2938758.
6. H. Yang et al., "Exploiting emotional concepts for image emotion recognition," *Vis. Comput.*, vol. 39, no. 5, pp. 2177–2190, 2023, doi: 10.1007/s00371-022-02472-8.
7. J. Zhang et al., "Image sentiment classification via multi-level sentiment region correlation analysis," *Neurocomputing*, vol. 469, pp. 221–233, Jan. 2022, doi: 10.1016/j.neucom.2021.10.062.
8. D. She et al., "WSCNet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1358–1371, May 2020, doi: 10.1109/TMM.2019.2939744.
9. X. Zhang et al., "DCNet: Weakly supervised saliency guided dual coding network for visual sentiment recognition," in *Proc. 26th Eur. Conf. Artif. Intell.*, 2023, pp. 3050–3057.
10. T. Yamamoto, S. Takeuchi, and A. Nakazawa, "Image emotion recognition using visual and semantic features reflecting emotional and similar objects," *IEICE Trans. Inf. Syst.*, vol. E104.D, no. 10, pp. 1691–1701, 2021, doi: 10.1587/transinf.2020EDP7218.
11. B. Zhou et al., "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.
12. J. Yang, D. She, and M. Sun, "Joint image emotion classification and distribution learning via deep convolutional neural network," in *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 3266–3272.
13. J. Yang, D. She, M. Sun, M.-M. Cheng, P. L. Rosin, and L. Wang, "Visual sentiment prediction based on automatic discovery of affective regions," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2513–2525, Sep. 2018, doi: 10.1109/TMM.2018.2803520.
14. H. Zhang and M. Xu, "Weakly supervised emotion intensity prediction for recognition of emotions in images," *IEEE Trans. Multimedia*, vol. 23, pp. 2033–2044, 2020, doi: 10.1109/TMM.2020.3007352.
15. S. Chen et al., "Reverse attention-based residual network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020, doi: 10.1109/TIP.2020.2965989.

XINYUE ZHANG is pursuing a Ph.D. degree at East China Normal University, Shanghai, 200062, China. 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, Singapore Management University, 178902, Singapore. She is the corresponding co-author of this article. Contact her at zxwang@smu.edu.sg.

GUITAO CAO is a professor at the Software Engineering Institute, East China Normal University, Shanghai, 200062, China. She is the corresponding co-author of this article. Contact her at gtcao@sei.ecnu.edu.cn.

SENG-BENG HO is currently CEO & Chief AI Scientist of AI Institute Global, 139955, Singapore. Contact him at sengbeng.ho@aiiglobal.ai.