Journal of Information Hiding and Multimedia Signal Processing ©2024 ISSN 2073-4212
Ubiquitous International Volume **15**, Number **4**, December **2024**

Multi-modal tooth decay recognition based on Contrastive Learning and Multi-label

Quynh Dao Thi Thuy*

Faculty of Information Technology
Posts and Telecommunications Institute of Technology (PTIT)
Intelligent Computing for Sustainable Development Laboratory (IC4SD, PTIT)
11398, Ha Noi, Viet Nam
quynhdtt@ptit.edu.vn

Cuong Nguyen Manh

Danang University of Medical Technology and Pharmacy
Da Nang, Viet Nam

Tien Nguyen Van

Faculty of Information Technology
Posts and Telecommunications Institute of Technology
11398, Ha Noi, Viet Nam

*Corresponding author: Quynh Dao Thi Thuy
Received July 23, 2024, revised October 29, 2024, accepted November 1, 2024.

Abstract. *In the field of image processing, multi-label learning poses a significant challenge, especially in cavity recognition. Additionally, combining multi-label learning with contrastive learning represents a substantial advancement in image processing and serves as a motivation to support cavity recognition in humans, aiding dentists in making more effective diagnoses. In this paper, we propose a deep learning model to learn both multi-label features and image features. Since multi-label features and image features are inherently discrete, using contrastive learning and a multi-label model is a way to connect them. Extensive experiments were conducted on the ImageNet-mini dataset and our self-collected P-Deltal dataset, achieving 85.65% and 89.28% mAP, respectively, for each dataset.*
Keywords: Multi-modal classiication, tooth decay recognition, Multi-label, Contrastive learning

1. **Introduction.** Dental caries is one of the most common oral diseases. Since humans constantly need to eat and drink, bacteria from food can ferment and create fissures on the surface of teeth, leading to cavities over the years. Dental caries is mostly observed in children, adolescents, and the elderly, but it can occasionally occur in adults and infants as well. Therefore, without proper dental care and early detection of cavities, it can significantly impact human health [1, 2].

Recently, artificial intelligence technology has advanced rapidly, particularly in image processing. This provides a foundation for us to build an AI model to identify images of dental caries, thereby assisting doctors in diagnosing and treating the condition quickly and effectively. However, most AI models require extensive data labeling for training, which is time-consuming and costly [3, 4]. Thus, in recent years, several research groups

[5, 6] have proposed more efficient image feature representation models to facilitate easier and more effective training.

Choudhury et al. [7] expanded the approach by using zero-shot learning mechanisms to learn objectives based on an image n-gram dictionary and predict the highest-scoring class. The research by Gao [8] and Dong [9] recently demonstrated the potential of using Transformer models to learn image features from textual features. However, the integration of learning image and text representations remains limited. Additionally, the accuracy of these technologies is relatively low because most zero-shot learning models rely on weak supervision with narrower guidance.

To address these issues, we propose a multi-model learning framework that connects both text and image modalities based on contrastive learning and multi-label properties (named MCLM) to learn shared representation spaces between text and image modalities through an optimized multi-label feature matrix. Specifically, we construct correlation matrices between text and images based on zero-shot learning mechanisms to solve the alignment problem between text, images, and labels. We aim to optimize this latent ranking matrix, where each row corresponds to images, each column corresponds to supporting text prompt captions, and furthermore, we integrate multi-label weights for each value in the matrix to rank label points corresponding to images and text. This allows us to focus attention on matching points and support the prediction of dental caries effectively.

Our contributions are three folds and are summarized as follows:

- **Novel methodology**: We propose a multi-modal framework combining image features and text features based on contrastive learning and multi-label techniques (MCLM) to support the identification of dental caries.
- **New dataset**: We introduce a uniquely collected dataset focused on RGB image data of dental caries, named P-Dental.
- **Analysis and evaluation**: We evaluate our multi-model approach in two ways: based on contrastive learning, multi-label learning, and multi-feature learning. We conduct evaluations on two datasets: one is a public dataset, and the other is a uniquely collected dataset.

The remainder of this paper is structured as follows. Section 2 discusses relevant previous studies. Section 3 presents our method. The experimental evaluation is shown in Section 4. Finally, some concluding remarks and a brief discussion are provided in Section 5.

2. Related works.

In this section, we will examine some notable works on multi-modal learning techniques, contrastive learning techniques, and multi-label techniques to support the identification of dental caries in images, followed by a brief introduction to modern works in the field of dental caries image recognition.

2.1. Multi-modal learning techniques.

The combination of text-based and image-based learning plays a crucial role in the field of image processing. Deep multi-modal models [10, 11] typically involve training with diverse datasets. Additionally, few-shot learning [12, 13] and zero-shot learning [14, 15] models have seen significant advancements. For instance, CLIP [16] and FILIP [17] process image-text pairs to learn joint representations. Numerous studies have demonstrated that combining multiple data sources, or multi-modal training, yields high efficiency, such as in multi-modal object detection [18] and multi-modal object segmentation [19]. In this research, we aim to develop a multi-modal learning approach for the problem of dental caries detection using limited image-text prompt pairs.

2.2. **Contrastive learning techniques.** Contrastive learning with the key idea that each class is assigned a target vector to represent features and learn in the most intuitive way [20, 21]. Most contrastive learning problems rely on few-shot learning mechanisms to optimize and find better target vectors [22, 23]. Alternative loss functions are often proposed to modify the reference label distribution in contrastive learning, such as Mixup [24], CutMix [25], or label smoothing [26]. Families of self-supervised representation models use contrastive learning [27, 28]. In some studies [29, 30], several loss functions based on metric learning have been proposed to learn robust representations. These loss functions are typically trained on some auxiliary knowledge such as images related to labels or frames from randomly selected videos, thus assuming that these approaches yield a very low false negative rate.

2.3. **Multi-label techniques.** Multi-label classification is designed to serve the purpose of categorizing with multiple class labels [31, 32]. Cheng et al. [33] employed GCN to build and predict multi-labels. Zhang et al. [34] used multi-labels to learn attribute features. The research team [35] incorporated Transformer into the multi-label problem to focus on the dependencies of feature sets. Most multi-label problems aim to extract as much label information as possible without resorting to conventional detection tasks, which is also one of the main advantages of the multi-label method.

2.4. **Discussion.** The aforementioned methods offer advantages in the process of training deep learning models. For instance, multi-model learning allows the model to understand various data sources, correlation learning enables the model to train on limited data, and multi-label learning helps us predict more labels and information. Therefore, we have combined these methods to research and develop solutions for the problem of tooth decay detection.

3. Material and methods.

3.1. **MCML framework.** The MCLM framework (illustrated in Figure 1) is proposed to support the recognition of dental caries in the human mouth. At the beginning of the diagram, the input is an image, and ResNet-101 is used as the backbone for feature extraction. The features extracted from the Multi-label model are predicted into labels and forwarded to the proposed loss function. The next step processes these features to distinguish them.

The steps for distinguishing features are as follows: transforming the features into feature space, placing them into the workspace according to the object's specific features through localization and grouping, and finally ensuring class separation by constructing class distance functions based on contrastive learning. We strive to build on multi-label classification technology to learn the most intuitive representations for dental caries objects. The proposed loss function evaluates the features from the network's multi-label classes and distinguishes them within the specific feature spaces of each class. This allows multi-label object analysis in the image to focus on key regions.

3.2. **Multi-label model.** Regarding the Multi-label model, we leverage the C-GCN model [36] to perform multi-label training for the problem of dental caries recognition in the human mouth. This model mainly comprises three blocks: the feature representation block, the multi-label learning block, and the multi-label score weighting block. The feature representation block uses the ResNet-101 model as the backbone to extract image features. The multi-label learning block is based on the GCN model [36], where we have a set of potential labels in the image to learn the graph and train it to find the optimal graph. The output of the multi-label learning block $\hat{Y}$ is a matrix that represents the

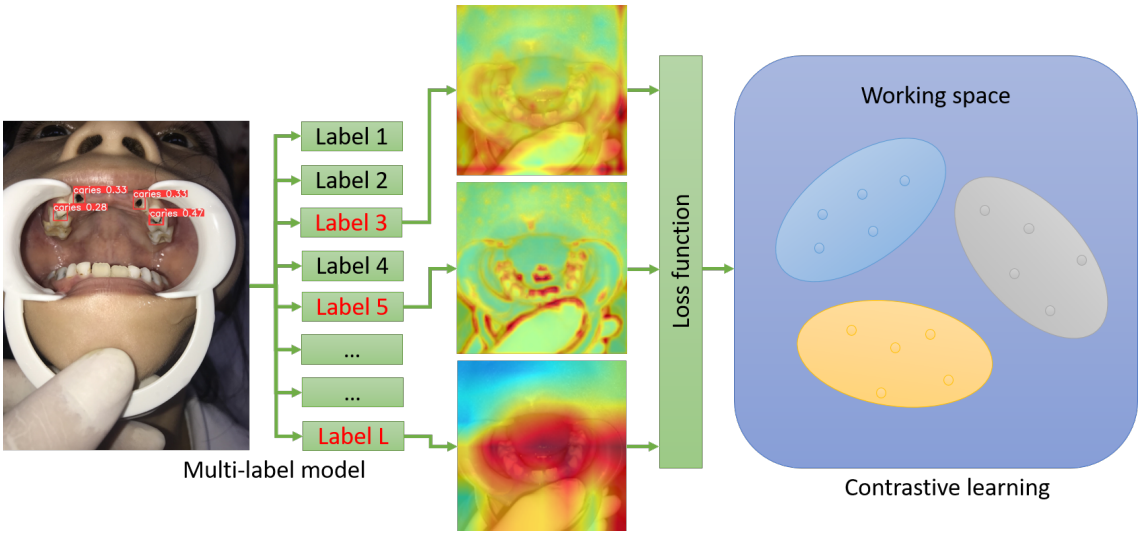

FIGURE 1. Our MCLM framework consists of two main components: Multi-label model and Contrastive learning

relationships between the labels and supports classification for the corresponding labels, also known as a correlation matrix. When this correlation matrix is multiplied by the feature representation matrix, it produces a set of multi-label vectors, representing the score weights of each label, meaning that labels likely to be present in the image will have increased weights. This is the multi-label score weighting block. Naturally, the Multi-label model also includes a loss function to evaluate the model's effectiveness, which we have named L_M. The formulas for $\hat{Y}$ are shown in equation (1).

$$\hat{Y} = Multi - label\ model(Image) \tag{1}$$

3.3. **Pre-processing Data.** In this section, we describe the method for processing distinct feature data from the multi-label classification model. The proposed loss function, L_P, is based on features from the network's multi-label classes and differentiates them using class-specific attention maps. We consider a deep neural network model comprising a feature extractor f and an object analyzer g. The function f, in conjunction with g, extracts the correct features for the objects from the multi-label classifications of the Multi-label model. A one-to-one and one-to-many mapping is constructed between f and g. At this stage, we have not implemented many-to-many mappings due to resource constraints and assume the model possesses all features of each class for prediction. This mapping simplifies the task of object analysis and allows the separation of image features according to multi-class representations. The formulas for f, g, and L_P are shown in equations (2), (3) and (4).

$$X = f(Image, \theta) \tag{2}$$

$$y = g(\hat{Y}, X, \beta) \tag{3}$$

$$\{\beta^*, \theta^*\} = argmin_{\theta,\beta}(L_P) \tag{4}$$

3.4. **Working Space.** To enhance the quality of dental caries recognition, we incorporate contrastive representation learning [21]. This method aims to improve the linkage between features and labels. Essentially, it involves constructing mappings from the feature space

TABLE 1. Detailed information about the number of images in the dataset P-Dental

Tooth decay labels	Number of images
Level 1	495
Level 2	510
Level 3	483
Level 4	498
Level 5	429

to another embedding space (the working space) to enhance accuracy and effectiveness. Transferring to an embedding space s_i helps reduce dimensionality while maintaining contrast and linkage between features within the same label y_j. Prediction in this context will revert to being based on the probability of an active label. The loss function used to evaluate the contrastive learning model in the working space is L_S, show in equation (5).

$$L_S = \sum_{j=1}^{L} \sum_{i=1}^{K} (y_j log s_i + (1 - y_j) log(1 - s_i)) \tag{5}$$

3.5. **Loss function.** To comprehensively evaluate the effectiveness of combining the models, we construct an overall loss function (also known as the final loss) to assess our entire model. Here, we have three loss functions: the multi-model loss, the data processing loss, and the contrastive learning loss. Each loss function corresponds to a specific model and task. The final loss will be the sum of these three loss functions, as shown in equation (6).

$$L_{final} = L_M + L_P + L_S \tag{6}$$

4. **Experiments.** In this section, we detail the implementation of the proposed model for detecting dental caries and the evaluation metrics on two datasets: ImageNet-mini [37] and our dataset, P-Dental. We also compare our model with state-of-the-art methods through quantitative and qualitative summary studies to analyze the superiority of the proposed model. Additionally, we discuss experiments conducted to evaluate the effectiveness of each module within the proposed model.

4.1. **Dataset.** In this work, two datasets are used for experimental evaluation: ImageNet-mini [37] and P-Dental (a dataset we collected ourselves). These datasets are introduced as follows.

ImageNet-mini contains 60,000 images with 100 different classes. Here, we only use 20,000 images with 10 classes to evaluate our model. We split the dataset into 60-20-20, with 60% for training, 20% for validation, and 20% for testing.

P-Dental is a dataset of dental caries collected from a dental clinic in Hanoi, Vietnam, with the permission of the Vietnamese ethics committee. The P-Dental dataset includes 2,415 images with 5 labels corresponding to different stages of dental caries. We collected these images using mobile devices to photograph the patients' oral cavities during dental examinations. Detailed information about the dataset is presented in Table 1.

4.2. **Performance metrics.** To evaluate the performance of the proposed model, we use the mean Average Precision (mAP) [38], which is a commonly used evaluation method for supporting multi-label classification and dental caries detection. And calculate AP, there are two methods: k-point interpolation and interpolation of all points.

TABLE 2. Comparisons with state-of-the-art methods on the MS-COCO and MLIC-Edu dataset.

Model	ImageNet-mini	P-Dental (our dataset)
CNN-RNN [43]	75.39	79.53
HCP [44]	76.74	81.62
RNN-Attention [45]	78.53	83.26
SSGRL [46]	80.68	85.92
BCE [47]	83.56	88.25
MCLM (ours)	**85.65**	**89.28**

$$AP = \frac{1}{k} \sum_{r} maxP(\tilde{r}) \tag{7}$$

where $P(\tilde{r})$ is the precision at recall $\tilde{r}$ with $\tilde{r} > r$.

4.3. **Training setup.** For our experiments, we use ResNet-101 [39] as the backbone for feature extraction from the images. We employ the ReLU activation function [40] for output results. During training, we utilize the Adam optimizer [41] with a learning rate of 0.0001. The batch size is set to 64. For each input, we resize images of varying sizes to 448x448, followed by data augmentation using PyTorch [42]. The label sets we use for each dataset are fixed; for example, P-Dental consistently uses 5 labels for which we train the GCN on these 5 labels.

4.4. **Experiment setup.** We design our extensive empirical study to answer the following three key research questions (RQs):

- RQ1: How is the MCLM model better compared to other deep learning methods with the same concept?
- RQ2: How does each situation in MCLM contribute to accurate deep learning?
- RQ3: How close is the prediction of the MCLM model to the ground truth?

In RQ1, we showcase the experiments conducted on the three foundational network baselines. For RQ2, we carried out a total of three distinct scenarios. Furthermore, for RQ3, we used the MCLM model to make predictions and provided some of the model's prediction results. The results will be averaged over experimental runs on two datasets.

4.5. **Results and discussion.**

4.5.1. *Comparison With Four Baselines (RQ1).* The comparison results with current methods under the same constructed scenario are presented in Table 2. It can be seen that representation learning through contrastive learning combined with a multi-label model yields superior results compared to basic methods and some of the latest modern methods. Moreover, the proposed method also significantly outperforms all methods in the previous rows. The MCLM model shows a relatively high improvement in average precision (mAP) by 2-3% compared to basic methods. This is achieved due to its ability to understand image features as well as multi-label characteristics, thereby integrating and enhancing efficiency and performance.

TABLE 3. The results of five experiments are presented.

Model	ImageNet-mini	P-Dental (our dataset)
MCLM (ours)	**85.65**	**89.28**
- Multi-label model	81.32	84.93
- branch Pre-processing Data	83.43	85.79
- Contrastive learning	82.56	84.23

4.5.2. *Applicability to Scenarios (RQ2).* The results comparing the effectiveness when adding modules and the feasibility of each module are shown in Table 3. Firstly, if the multi-label model module is removed, resulting in the model lacking multi-label features, the accuracy decreases by 4-5%, which clearly indicates the importance of multi-label features. Secondly, if the data processing step from the multi-label stage is omitted, the model's accuracy also decreases, as this process helps the model understand multi-labels and enhances the features supporting the recognition process. Lastly, removing the contrastive learning module significantly reduces the model's accuracy because this step supports the efficient linking of features through the Working Space and enhances the model's recognition capability.

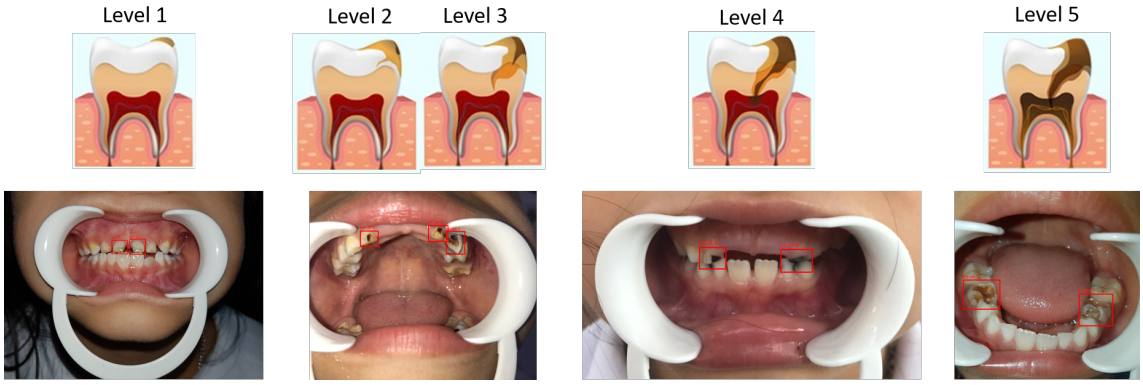

FIGURE 2. Predictions of the proposed method (MCLM).

4.5.3. *Qualitative study (RQ3).* We conducted experiments on the test set to visualize the model's prediction results on the P-Deltal set, as shown in Figure 2. The first row represents a symbolic example with the associated cavity severity labels. The second row shows the prediction results for each label, aligned with the columns of the symbolic example above. From this visualization, it is evident that the model performs quite well. Although the difficulty of this classification task is fairly high, the ability to learn based on features and multi-labels has supported the model's accurate predictions. The confusion between cavities at level 2 and level 3 is minimal but still constitutes a significant percentage. Meanwhile, levels 1, 4, and 5 are clearly distinct in terms of features, resulting in almost no prediction errors. Overall, the proposed model achieves stable performance.

5. **Conclusions.** In summary, to support the identification of cavities in humans, challenges such as low-resolution images and the difficulty of recognizing cavities in the oral cavity have motivated us to improve image recognition methods. Leveraging the advantages of multi-label and contrastive learning, we have skillfully combined them to enhance the accuracy of cavity identification. Building contrastive learning frameworks based on multi-labels helps effectively map image features to label features. We also integrated the loss functions of each module to comprehensively improve accuracy. Our proposed MCLM

model is advanced, demonstrating superiority in experimental results. Additionally, we collected data on cavities named P-Deltal. However, the MCLM model has drawbacks in prediction latency and performs well with small datasets. Therefore, we plan to improve and overcome these two drawbacks in future research.

Acknowledgement. This research is funded by the Posts and Telecommunications Institute of Technology (PTIT), Vietnam under grant number "13-2024-HV-CNTT1". The authors would like to thank PTIT for the financial support.

REFERENCES

[1] Ren, Y. F., Rasubala, L., Malmstrom, H., & Eliav, E. (2020). Dental care and oral health under the clouds of COVID-19. JDR Clinical & Translational Research, 5(3), 202-210.
[2] Li, Y., Tang, H., Liu, Y., Qiao, Y., Xia, H., & Zhou, J. (2022). Oral wearable sensors: health management based on the oral cavity. Biosensors and Bioelectronics: X, 10, 100135.
[3] Zha, D., Bhat, Z. P., Lai, K. H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). Data-centric artificial intelligence: A survey. arXiv preprint arXiv:2303.10158.
[4] Adadi, A. (2021). A survey on data-efficient algorithms in big data era. Journal of Big Data, 8(1), 24.
[5] Guo, C., Fan, B., Zhang, Q., Xiang, S., & Pan, C. (2020). Augfpn: Improving multi-scale feature learning for object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 12595-12604).
[6] Yuan, X., Lin, Z., Kuen, J., Zhang, J., Wang, Y., Maire, M., ... & Faieta, B. (2021). Multimodal contrastive training for visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6995-7004).
[7] Choudhury, S., Laina, I., Rupprecht, C., & Vedaldi, A. (2024). The curious layperson: Fine-grained image recognition without expert labels. International Journal of Computer Vision, 132(2), 537-554.
[8] Gao, C., Cai, G., Jiang, X., Zheng, F., Zhang, J., Gong, Y., ... & Bai, X. (2022). Conditional feature learning based transformer for text-based person search. IEEE Transactions on Image Processing, 31, 6097-6108.
[9] Dong, X., Zhang, H., Zhu, L., Nie, L., & Liu, L. (2022). Hierarchical feature aggregation based on transformer for image-text matching. IEEE Transactions on Circuits and Systems for Video Technology, 32(9), 6437-6447.
[10] Wang, Y. (2021). Survey on deep multi-modal data analytics: Collaboration, rivalry, and fusion. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(1s), 1-25.
[11] Gupta, D., Suman, S., & Ekbal, A. (2021). Hierarchical deep multi-modal network for medical visual question answering. Expert Systems with Applications, 164, 113993.
[12] Baik, S., Choi, J., Kim, H., Cho, D., Min, J., & Lee, K. M. (2021). Meta-learning with task-adaptive loss function for few-shot learning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9465-9474).
[13] Simon, C., Koniusz, P., Nock, R., & Harandi, M. (2020). Adaptive subspaces for few-shot learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4136-4145).
[14] Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., ... & Wu, Q. J. (2022). A review of generalized zero-shot learning methods. IEEE transactions on pattern analysis and machine intelligence, 45(4), 4051-4070.
[15] Xu, W., Xian, Y., Wang, J., Schiele, B., & Akata, Z. (2020). Attribute prototype network for zero-shot learning. Advances in Neural Information Processing Systems, 33, 21969-21980.
[16] Fan, L., Krishnan, D., Isola, P., Katabi, D., & Tian, Y. (2024). Improving clip training with language rewrites. Advances in Neural Information Processing Systems, 36.
[17] Huang, R., Long, Y., Han, J., Xu, H., Liang, X., Xu, C., & Liang, X. (2023, June). Nlip: Noise-robust language-image pre-training. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 926-934).
[18] Wang, L., Zhang, X., Song, Z., Bi, J., Zhang, G., Wei, H., ... & Zhao, L. (2023). Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. IEEE Transactions on Intelligent Vehicles, 8(7), 3781-3798.

[19] Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., ... & Dietmayer, K. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Transactions on Intelligent Transportation Systems, 22(3), 1341-1360.

[20] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... & Krishnan, D. (2020). Supervised contrastive learning. Advances in neural information processing systems, 33, 18661-18673.

[21] Le-Khac, P. H., Healy, G., & Smeaton, A. F. (2020). Contrastive representation learning: A framework and review. Ieee Access, 8, 193907-193934.

[22] Liu, C., Fu, Y., Xu, C., Yang, S., Li, J., Wang, C., & Zhang, L. (2021, May). Learning a few-shot embedding model with contrastive learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 10, pp. 8635-8643).

[23] Zeng, Q., & Geng, J. (2022). Task-specific contrastive learning for few-shot remote sensing image scene classification. ISPRS Journal of Photogrammetry and Remote Sensing, 191, 143-154.

[24] Chou, H. P., Chang, S. C., Pan, J. Y., Wei, W., & Juan, D. C. (2020). Remix: rebalanced mixup. In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16 (pp. 95-110). Springer International Publishing.

[25] Pan, H., Guo, Y., Yu, M., & Chen, J. (2024). Enhanced Long-Tailed Recognition with Contrastive CutMix Augmentation. IEEE Transactions on Image Processing.

[26] Liang, J., Li, L., Bing, Z., Zhao, B., Tang, Y., Lin, B., & Fan, H. (2022, October). Efficient one pass self-distillation with zipf's label smoothing. In European conference on computer vision (pp. 104-119). Cham: Springer Nature Switzerland.

[27] Kotar, K., Ilharco, G., Schmidt, L., Ehsani, K., & Mottaghi, R. (2021). Contrasting contrastive self-supervised representation learning pipelines. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 9949-9959).

[28] Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., & Tang, J. (2021). Self-supervised learning: Generative or contrastive. IEEE transactions on knowledge and data engineering, 35(1), 857-876.

[29] Jiang, W., Huang, K., Geng, J., & Deng, X. (2020). Multi-scale metric learning for few-shot learning. IEEE Transactions on Circuits and Systems for Video Technology, 31(3), 1091-1102.

[30] Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., & Cohen, J. P. (2020, November). Revisiting training strategies and generalization performance in deep metric learning. In International Conference on Machine Learning (pp. 8242-8252). PMLR.

[31] Liu, W., Wang, H., Shen, X., & Tsang, I. W. (2021). The emerging trends of multi-label learning. IEEE transactions on pattern analysis and machine intelligence, 44(11), 7955-7974.

[32] Law, A., & Ghosh, A. (2021). Multi-label classification using binary tree of classifiers. IEEE Transactions on Emerging Topics in Computational Intelligence, 6(3), 677-689.

[33] Cheng, Y., Ma, M., Li, X., & Zhou, Y. (2021). Multi-label classification of fundus images based on graph convolutional network. BMC Medical Informatics and Decision Making, 21, 1-9.

[34] Zhang, S., Xu, R., Xiong, C., & Ramaiah, C. (2022). Use all the labels: A hierarchical multi-label contrastive learning framework. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16660-16669).

[35] Rodríguez, M. A., AlMarzouqi, H., & Liatsis, P. (2022). Multi-label retinal disease classification using transformers. IEEE Journal of Biomedical and Health Informatics, 27(6), 2739-2750.

[36] Nie, W., Ren, M., Nie, J., & Zhao, S. (2020). C-GCN: Correlation based graph convolutional network for audio-video emotion recognition. IEEE Transactions on Multimedia, 23, 3793-3804.

[37] Yu, H., Cheng, X., & Peng, W. Supplementary Materials of TOPLight: Lightweight Neural Networks with Task-Oriented Pretraining for Visible-Infrared Recognition.

[38] Li, C., Yang, T., Zhu, S., Chen, C., & Guan, S. (2020). Density map guided object detection in aerial images. In proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 190-191).

[39] Vaishali, S., & Neetu, S. (2024). Enhanced copy-move forgery detection using deep convolutional neural network (DCNN) employing the ResNet-101 transfer learning model. Multimedia Tools and Applications, 83(4), 10839-10863.

[40] Alhassan, A. M., & Zainon, W. M. N. W. (2021). Brain tumor classification in magnetic resonance image using hard swish-based RELU activation function-convolutional neural network. Neural Computing and Applications, 33(15), 9075-9087.

[41] Swathi, T., Kasiviswanath, N., & Rao, A. A. (2022). An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. Applied Intelligence, 52(12), 13675-13688.

[42] Sinha, A., Ayush, K., Song, J., Uzkent, B., Jin, H., & Ermon, S. (2021). Negative data augmentation. arXiv preprint arXiv:2102.05113.

[43] Zhou, X., Li, Y., & Liang, W. (2020). CNN-RNN based intelligent recommendation for online medical pre-diagnosis support. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 18(3), 912-921.

[44] Tang, H., Ma, G., Guo, L., Fu, X., Huang, H., & Zhan, L. (2022). Contrastive brain network learning via hierarchical signed graph pooling model. IEEE transactions on neural networks and learning systems.

[45] Gu, C., Wang, S., Zhu, Y., Huang, Z., & Chen, K. (2021, January). Weakly supervised attention rectification for scene text recognition. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 779-786). IEEE.

[46] Pu, T., Chen, T., Wu, H., & Lin, L. (2022, June). Semantic-aware representation blending for multi-label image recognition with partial labels. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2091-2098).

[47] Xu, Z., Liu, R., Yang, S., Chai, Z., & Yuan, C. (2023). Learning imbalanced data with vision transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 15793-15803).