

Category-Specific Selective Feature Enhancement for Long-Tailed Multi-Label Image Classification

Ruiqi Du, Xu Tang*, Xiangrong Zhang, Jingjing Ma
Xidian University, China

24171111300@stu.xidian.edu.cn, tangxu128@gmail.com,
xrzhang@mail.xidian.edu.cn, jjma@xidian.edu.cn

Abstract

Since real-world multi-label data often exhibit significant label imbalance, long-tailed multi-label image classification has emerged as a prominent research area in computer vision. Traditionally, it is considered that deep neural networks' classifiers are vulnerable to long-tailed distributions, whereas the feature extraction backbone remains relatively robust. However, our analysis from the feature learning perspective reveals that the backbone struggles to maintain high sensitivity to sample-scarce categories but retains the ability to localize specific areas effectively. Based on this observation, we propose a new model for long-tailed multi-label image classification named category-specific selective feature enhancement (CSSFE). First, it utilizes the retained localization capability of the backbone to capture label-dependent class activation maps. Then, a progressive attention enhancement mechanism, updating from head to medium to tail categories, is introduced to address the low-confidence issue in medium and tail categories. Finally, visual features are extracted according to the optimized class activation maps and combined with semantic information to perform the classification task. Extensive experiments on two benchmark datasets highlight our findings' generalizability and the proposed CSSFE's superior performance. Our code is available at <https://github.com/TangXu-Group/multilabelRSSC/tree/main/CSSFE>.

1. Introduction

With the advancement of deep learning techniques, the performance of multi-label image classification has been enhanced significantly [6, 7, 46]. However, due to the inherent imbalance in the occurrence frequency of various objects [22, 28, 30], the label distribution of real-world multi-label image datasets typically exhibits the long-tailed phenomenon [14, 15, 32]. In other words, a minority of la-

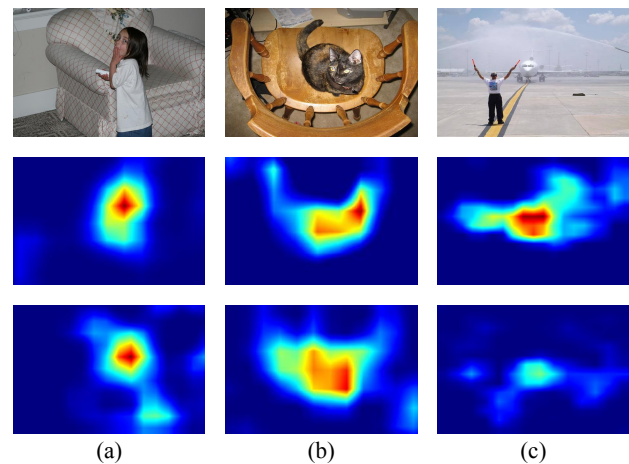


Figure 1. We set three visual comparison groups: (a) Head category: person; (b) Medium category: cat; (c) Tail category: airplane. The model in the second row was trained on a traditionally label-balanced multi-label dataset, while the model in the third row was trained on a long-tailed distribution multi-label dataset. Comparative analysis of class activation maps visualizations between the second row and the third row (CAM_{lb} vs. CAM_{lt}), we observe that the primary impact of long-tailed multi-label distribution on deep neural networks manifests at the confidence level. Specifically, CAM_{lb} and CAM_{lt} exhibit structural similarity in overall activation areas, while CAM_{lt} demonstrates significant attention weight attenuation especially for tail categories.

els appears frequently, while the majority has only limited sample data. This imbalanced distribution poses substantial challenges to conventional multi-label image classification models, as they tend to overfit head classes while neglecting the learning of tail classes [43]. To address these problems, long-tailed multi-label image classification [24, 37, 38] has emerged and attracted considerable attention.

There are two baseline strategies to address the head-to-tail imbalance problem in long-tailed distributions. One is to re-sample the samples in the datasets during train-

*:Corresponding author

ing to ensure that each class contributes to gradient propagation with similar frequency [9, 36, 41, 43]. Another is re-weighting the contributions of head and tail categories, allowing the model to prioritize less frequent tail samples while appropriately fading the influence of frequently learned head samples [16, 20, 34]. However, in the long-tailed multi-label image classification task, the effectiveness of the above methods cannot meet what we expected due to the symbiotic relationships between labels [18, 33]. In addition, inspired by the long-tailed single-label image classification methods, existing studies generally believe that the main impact of long-tailed multi-label distribution on deep neural networks is the overfitting of the classifier to the head categories and the underfitting of the tail categories, while the performance of the feature extraction backbone is almost unaffected [40, 43]. Nevertheless, when we deeply investigate it from the perspective of feature learning, we find that this view is not entirely applicable to the long-tailed multi-label image classification task.

Specifically, we conducted control experiments by training two architecturally identical yet parameter-independent deep neural networks (see Section 3.2). One is trained on a long-tailed multi-label dataset, and the other is trained on a conventionally label-balanced multi-label dataset. Then, we introduced the class activation mapping (CAM) [44] technique to analyze their spatial attention distributions for different labels. The normalized visualization examples (as shown in Figure 1) and numerical evidence (see Section 3.2) reveal that although the model trained with long-tailed data manifests substantial attention attenuation toward tail categories compared to another model, it preserves fundamental capability in localizing category-specific discriminative regions.

Based on the above discussion, we conclude that the primary factor limiting the performance of models trained on long-tailed multi-label image datasets is the insufficient feature representation for the regions corresponding to the sample-scarce labels. To address this issue, this paper proposes a category-specific selective feature enhancement (CSSFE) model consisting of three parts: label-dependent attention region learning, progressive attention enhancement, and classification. They are responsible for identifying category-specific regions from the input image, selectively adjusting the model’s attention weights for underrepresented categories (i.e., medium and tail categories) using semantic relationships, and integrating visual features with semantic embeddings to complete the classification task, respectively.

The core contributions of our work are summarized as follows.

1. From the perspective of feature learning, we analyze the impact of long-tailed multi-label distribution on deep neural networks and find that the primary limitation hin-

dering model performance is the inadequate representation capability of the feature extraction backbone for partial categories. Fortunately, the fundamental ability of the feature extraction backbone to identify discriminative regions of specific categories is not sensitive to the head-tail imbalance problem.

2. Building on the above findings, we introduce a long-tailed multi-label image classification model (i.e., CSSFE). Through the one-way semantic connection-based attention adjustment strategy that flows from head to middle to tail categories, the model’s sensitivity to scarce categories such as the medium and tail can be successfully improved. Meanwhile, the typical trade-off where sacrificing the performance of the head categories in exchange for improving the performance of the tail categories can be avoided.
3. As a consequence, CSSFE achieves state-of-the-art results on two benchmark long-tailed multi-label image classification datasets. Extensive experiments not only validate the performance of our CSSFE but also demonstrate the rationality of our findings.

2. Related Work

In this section, we first review several classic models for long-tailed multi-label image classification, followed by methods for enhancing model interpretability.

One approach to addressing long-tailed multi-label image classification is the re-sampling strategy. For instance, Wu et al. [36] employed a category-aware re-sampling method, commonly used in long-tailed single-label classification, to increase the likelihood of learning tail labels. While effective to some extent, the inherent label co-occurrence characteristics in multi-label data still lead to a significant class imbalance in the re-sampled training set. To mitigate this issue, Chen et al. [3] proposed a group sampling strategy that ensures a more balanced data distribution within sub-datasets. Meanwhile, Guo et al. [10] introduced a collaborative sampling approach that combines uniform sampling with rebalanced sampling to enhance the diversity of the data set. Another strategy is re-weighting. For example, distribution-balanced (DB) loss [36] and probability-guided (PG) loss [20] help prevent the model from overfitting the head classes while neglecting the tail classes by reducing the gradient update speed disparity between them. In addition, Yan et al. [39] combined the re-weighting strategy with a large visual language model, which not only maintained the category balance advantage of the re-weighting strategy but also used the strong representation ability of the pre-trained model to correct the long-tailed data distribution bias.

To enhance the interpretability of deep neural networks, it is feasible to visualize attention areas, which have been proven to be effective in fields such as object tracking,

change detection, and multi-modal retrieval [13, 31, 35]. Especially for label-related image classification tasks, visualizing areas associated with labels provides an intuitive way to observe the impact of label issues. For example, Zhang et al. [42] demonstrated that noisy labels lead to deviations in the regions where the model focuses. Therefore, they counteracted the influence of noisy labels by maintaining attention consistency. Similarly, in the context of partially annotated multi-label image classification, Kim et al. [17] investigated the difference between partial-label training and full-label training, intending to boost the model’s feature learning ability in partial-label scenarios. Inspired by the above approaches, this paper aims to analyze the differences in feature learning of head, medium, and tail categories to explain the impact of long-tailed multi-label on deep neural networks.

3. Proposed Method

3.1. Overview

The proposed CSSFE stems from our finding that deep neural networks maintain their capacity to locate label-associated regions yet exhibit diminished discriminative sensitivity toward tail categories under the influence of long-tailed label distributions (see Figure 1). To prove the reliability and generalizability of our motivation, we perform numerical comparison and gradient analysis in Section 3.2. Then, in Section 3.3 and Section 3.4, we introduce the detailed pipeline and optimization objective of CSSFE, whose framework is shown in Figure 2.

3.2. Feature Learning Analysis

In this section, we seek to explain the reasons for the performance limitations of deep neural networks facing long-tailed multi-label datasets through the aspect of feature learning. To this end, we devise a comparative framework to analyze the similarities and differences between the feature attention patterns of a model trained on a long-tailed multi-label dataset versus that trained on a label-balanced multi-label dataset. In detail, given an input image I , a backbone is used to generate its global visual feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, where H means the height, W refers to the width, and C is the number of channels. Subsequently, \mathbf{F} processes through a 2-D convolutional layer and yields class attention maps $\mathbf{CAM} \in \mathbb{R}^{H \times W \times c}$, where c corresponds to the number of categories. The above process is depicted in the label-dependent attention region learning of Figure 2. Then, the prediction logit s_i for category i can be computed via the weighted sum of the i -th class attention map and participates in the training process. Therefore, CAM can effectively reflect the label-associated attention regions [44].

We conducted systematic experiments using the above

model on two distinct datasets: the label-balanced COCO dataset [21] and its long-tailed variant LT-COCO [36]. Through evaluation on the test set of LT-COCO, we counted the class activation maps generated by the two models, denoted as \mathbf{CAM}_{lb} (label-balanced COCO training) and \mathbf{CAM}_{lt} (long-tailed LT-COCO training). To prove the structural preservation of attention distribution, we first computed Spearman correlation coefficients between \mathbf{CAM}_{lb} and \mathbf{CAM}_{lt} as the experiment group, which can effectively reflect the consistency of the overall structure of the class activation maps through the spatial attribution rankings of the corresponding pixels [17]. In addition, we simulated random attention distribution by counting the 2-D Gaussian images with \mathbf{CAM}_{lt} as the control group. Figure 3a presents the comparative results, indicating two critical observations. First, the experimental groups of different categories all exhibit strong positive correlations (approaching unity). Second, the control group demonstrates near-zero relevance, confirming the non-randomness of learned attention maps. This discrepancy implies that the model’s ability to focus on label-related regions is maintained despite long-tailed distribution challenges.

To further investigate the attention attenuation phenomenon induced by long-tailed multi-label distributions, we perform a comparative analysis of attention scores across different classes between \mathbf{CAM}_{lb} and \mathbf{CAM}_{lt} . Note that our analysis focuses on the top 50% quantity of attention values to mitigate noise interference from low-attention regions. As shown in Figure 3b, it can be found that as the category frequency decreases, the reliability of attention location areas gradually decreases. Especially for the tail samples, the overall \mathbf{CAM}_{lt} has a significant attenuation compared to \mathbf{CAM}_{lb} . This finding reveals long-tailed multi-label distributions impair the model’s sensitivity to tail category-dependent features, which can be further theoretically justified via gradient analysis. In the long-tailed multi-label image classification task, binary cross-entropy loss and its variants [12] are widely used. For robustness, we also analyze the gradient of the original binary cross-entropy loss [27] \mathcal{L}_{bce} , which can be formulated as,

$$\mathcal{L}_{bce} = - \sum_{i=1}^c [y_i \log \sigma(s_i) + (1 - y_i) \log(1 - \sigma(s_i))] \quad (1)$$

where y_i is the label of i -th category, $\sigma(\cdot)$ refers to the sigmoid function, and s_i is the classification logit of i -th category. The gradient of \mathcal{L}_{bce} with respect to s_i is $\frac{\partial \mathcal{L}_{bce}}{\partial s_i} = \sigma(s_i) - y_i$. Due to the scarcity of positive samples, the model receives the gradient in most cases as $\sigma(s_i) - 0 > 0$ for tail categories. This positive gradient results in a systematic suppression of the tail category logits. Since the logit values are equivalent to the mean of the class activation map, and the class activation map reflects the degree of

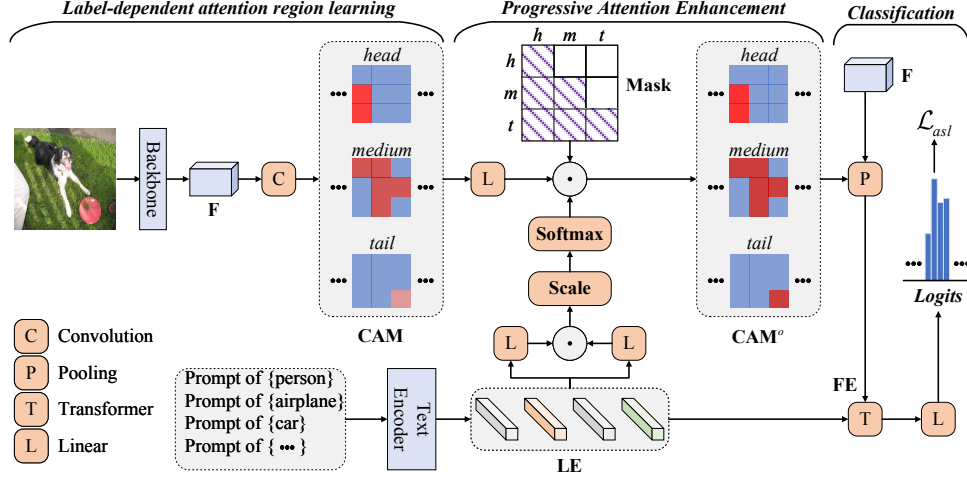


Figure 2. The overall framework of our proposed long-tailed multi-label image classification methods. It comprises three sub-parts: label-dependent attention region learning, progressive attention enhancement, and classification. In label-dependent attention region learning, the relevant attention areas of different categories are located according to the visual features extracted by the backbone. The progressive attention enhancement module exploits label correlation to adjust the weights of class activation maps according to their characteristics, i.e., the feature learning status of head categories is reliable, the middle categories follow, and the tail categories have obvious attention decay. Finally, the visual embeddings located according to the optimized class activation maps and semantic embeddings extracted by the language model are interacted with and used for classification.

attention paid by the model to the label-related area, it ultimately makes it difficult for the model to determine whether the tail categories exist in the image.

3.3. Progressive Attention Enhancement Module

At this point, we can confidently conclude that deep neural networks maintain effective region localization capabilities in long-tailed multi-label image classification tasks. However, reliably generating attention weights still remains a challenge. Therefore, improving the model’s focus on relevant regions associated with low-frequency categories is the key to achieving high-precision classification. Considering the complex symbiotic relationship among various objects, we hope to propose a class activation map optimization mechanism grounded in label relevance to assist deep neural networks in better capturing label-associated regions. To this end, a progressive attention enhancement module is proposed in this section.

Given the candidate label text set {person, airplane, dog, ...}, we first encode them into label embeddings $\mathbf{LE} \in \mathbb{R}^{c \times d}$, where d denotes the dimension of label embeddings. This encoding is achieved using the text encoder of the pre-trained CLIP model [25] with standardized prompt templates (e.g., “a photo of {class}”). Note that the text encoder does not participate in the training process, but the label embeddings undergo a single linear transformation layer to facilitate cross-dataset knowledge transfer. Next, the transferred label embeddings are projected into the query and key vectors through independent linear layers. Subse-

quently, they are processed through dot-product attention, scaled scaling, and softmax normalization to generate the label co-occurrence attention matrix $\mathbf{M} \in \mathbb{R}^{c \times c}$.

As experimentally validated in Section 3.2, the reliability of label-associated attention regions exhibits a statistically significant positive correlation with category volumes. This empirical evidence confirms the capability of deep neural networks in capturing discriminative features for head categories while highlighting the progressive degradation of confidence from medium to tail categories. To mitigate this problem, we introduce a distribution-aware binary mask $\mathbf{Mask} \in \mathbb{R}^{c \times c}$, which enforces a one-way attention flow from head to medium to tail categories while blocking reverse interactions, thereby selectively amplifying the semantic weights of the middle and tail categories. Finally, class activation maps \mathbf{CAM} are linearly mapped and multiplied by the label co-occurrence attention matrix and the binary mask to obtain the final optimized class activation map \mathbf{CAM}^o .

The overall process discussed above can be formalized as follows,

$$\mathbf{M} = \delta \left(\frac{\Phi_Q(\mathbf{LE}) \odot \Phi_K(\mathbf{LE})^T}{\sqrt{d_k}} \right), \quad (2)$$

$$\mathbf{CAM}^o = \mathbf{M} \odot \Phi_V(\mathbf{CAM}) \odot \mathbf{Mask},$$

where $\delta(\cdot)$ denotes the softmax function, $\Phi_Q(\cdot)$, $\Phi_K(\cdot)$, $\Phi_V(\cdot)$ indicate linear layers, $\sqrt{d_k}$ is a scaling factor, and \odot illustrates element-wise multiplication operation. Let $\mathbf{M}_{a,b}$

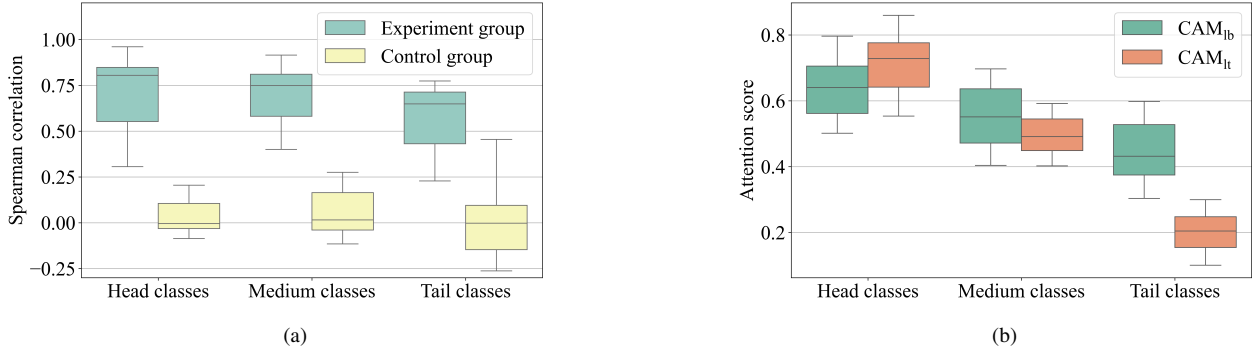


Figure 3. (a) **Overall Structural Analysis.** The experimental group calculates the Spearman correlation coefficient between CAM_{lb} and CAM_{lt} , while the control group involves the Spearman correlation coefficient between CAM_{lt} and the Gaussian-distributed image. The statistical results show that the correlation coefficient of the experimental group is significantly higher than that of the control group. This finding verifies that deep neural networks can still maintain a stable class activation map positioning ability under long-tailed multi-label distribution. (b) **Attention Weight Analysis.** By comparing the attention weight differences between CAM_{lb} and CAM_{lt} for head, middle, and tail categories, it is evident that as the number of samples decreases, the model’s attention in category-relevant regions decreases. Therefore, improving the model’s confidence in a specific category is the key to long-tailed multi-label image classification.

signify the semantic contribution weight from category of type b to category of type a , where $a, b \in \{h, m, t\}$ represent head, medium, and tail categories, respectively. $\mathbf{M}_{h,m}$, $\mathbf{M}_{h,t}$, $\mathbf{M}_{m,t}$ are set to 0, while all other elements remain 1. This directional isolation mechanism guides the model in prioritizing learning the feature representations of under-represented medium and tail categories to improve their semantic discriminability. Meanwhile, it prevents redundant information from negatively impacting the already well-represented class activation maps, thus avoiding the phenomenon where the performance of tail categories improves while the performance of head categories declines.

3.4. Classification and Optimization

When we get the optimized class activation maps, the class-specific visual features can be achieved. Then, the multi-label image classification tasks can be accomplished by integrating the semantic information from the label embeddings. Specifically, the visual feature map is multiplied with each channel of the CAM^o to generate distinct class-specific visual feature vectors \mathbf{FE} . These visual feature vectors are then concatenated with the corresponding semantic embeddings and processed through a Transformer block to facilitate interaction between visual and semantic information. Then, the fused representations are passed through a linear layer to produce the classification logits. Finally, the asymmetric loss \mathcal{L}_{asl} is adopted as our loss function, which is formulated as,

$$\mathcal{L}_{asl} = - \sum_{i=1}^c \left[y_i (1 - p_i)^{\gamma_+} \log(p_i) + (1 - y_i) p_i^{\gamma_-} \log(1 - p_i) \right], \quad (3)$$

where $p_i = \sigma(s_i)$ implies the prediction score for class i , and γ_+ and γ_- are regulating factors that are set as 0 and 4 in this paper, respectively. The reasons for choosing this loss are discussed in the Supplementary Material.

4. Experiments

4.1. Dataset Introduction

To evaluate the effectiveness of CSSFE, we select two benchmark datasets designed explicitly for long-tailed multi-label image recognition, including LT-VOC [36] and LT-COCO [36]. LT-VOC, a long-tailed variant of the PASCAL VOC dataset [8], contains 6,904 annotated images across 20 categories, with the number of samples per class ranging from 775 to 4. With 6,909 images, LT-COCO spans 80 object categories, where the maximum per-class sample size is 1,128 images, and the minimum contains only six images.

4.2. Implementation Details.

All experiments are conducted using the PyTorch framework on an NVIDIA 3090 GPU with 24GB of memory. The reported numerical scores are the average results from multiple random experiments. Following [36], LT-VOC is divided into 1,142 training images and 4,952 test images, while LT-COCO consists of 1,909 training images and 5,000 images for evaluation. Two backbones in CSSFE, i.e., ResNet50 [11] and CLIP [25], are initialized using the pre-trained weights. Other parts of CSSFE are initialized randomly. In the training stage, the Adam optimizer with a weight decay of 1×10^{-4} is selected to update the parameters. The learning rates for LT-VOC and LT-COCO are set

Table 1. Performance of different methods on LT-VOC and LT-COCO. We report the mAP (%) values for the total, head, medium, and tail classes, respectively. The best values are highlighted in bold.

Datasets	LT-VOC				LT-COCO			
Methods	Total	Head	Medium	Tail	Total	Head	Medium	Tail
ERM [36]	70.86	68.91	80.20	65.31	41.27	48.48	49.06	24.25
RW [36]	74.70	67.58	82.81	73.96	42.27	48.62	45.80	32.02
OLTR [23]	71.02	70.31	79.80	64.95	45.83	47.45	50.63	38.05
LDAM [1]	70.73	68.73	80.38	69.09	40.53	48.77	48.38	22.92
CB Focal [5]	75.24	70.30	83.53	72.74	49.06	47.91	53.01	44.85
BBN [45]	73.37	71.31	81.76	68.62	50.00	49.79	53.99	44.94
ASL [26]	76.40	70.70	82.26	76.29	50.21	49.05	53.65	46.68
DB Focal [36]	78.94	73.22	84.18	79.30	53.55	51.13	57.05	51.06
LTML [10]	81.44	75.68	85.53	82.69	56.90	54.13	60.59	54.47
CDRS+AFL [29]	78.96	73.35	85.03	78.63	55.35	52.45	59.48	52.46
Bilateral-TFS [19]	81.58	75.88	84.11	83.95	56.38	55.93	58.26	54.29
PG Loss [20]	80.37	73.67	83.83	82.88	54.43	51.23	57.42	53.40
COMIC [41]	81.53	73.10	89.18	84.53	55.08	49.21	60.08	55.36
CAE-Net [2]	81.61	74.00	85.35	85.28	57.64	52.37	61.18	57.63
CPRFL [39]	86.28	81.84	90.51	86.43	66.69	66.35	70.99	61.33
CSSFE (Ours)	86.38	81.93	91.16	87.26	66.93	66.96	71.21	62.41

as 5×10^{-5} and 1×10^{-5} , and the epochs are fixed as 80 empirically. In addition, the auto augmentation proposed in [4] is applied to stabilize the training. For the evaluation metric, we categorize the classes into head categories (with more than 100 samples), medium categories (with 20 to 100 samples), and tail categories (with fewer than 20 samples) [36]. Then, the mean average precision (mAP) is used to assess the model’s overall performance and its performance across various categories.

4.3. Comparisons with State-of-the-Art Methods

Two groups of compared methods are chosen to demonstrate the effectiveness of our proposed CSSFE. The first group includes classic methods for long-tailed recognition, such as ERM [36], RW [36], OLTR [23], LDAM [1], CB Focal [5], BBN [45], and ASL [26]. The second group consists of algorithms designed explicitly for long-tailed multi-label image classification tasks, including DB Focal [36], LTML [10], CDRS+AFL [29], Bilateral-TFS [19], PG Loss [20], COMIC [41], CAE-Net [2], and CPRFL [39]. It is worth noting that CPRFL shares the same text pre-trained parameters as our text encoder for fairness. Detailed descriptions of most comparison methods can be found in Section 2.

Based on the performance results summarized in Table 1, it is easy to find that CSSFE outperforms other methods in all cases. Specifically, conventional long-tailed classification methods maintain relatively acceptable performance in the head and medium categories but show notable weakness in the tail classes. For example, comparing BBN with DB Focal horizontally, one is a well-performing method in the first group, while the other is

relatively weaker in the second group. We can find their MAP scores for the head categories on LT-COCO differ by only 1.34%, but for the tail categories, the difference reaches 6.12%. This highlights that the traditional re-sampling or re-weighting paradigms struggle to address the label coupling issues in the multi-label scenario. Additionally, the performance of some specifically designed models fluctuates across different the class types. For instance, COMIC outperforms PG Loss in the medium and tail classes, but its performance in head classes is weaker. Ultimately, our CSSFE performs best across all metrics. Compared to the second-best method, CPRFL, CSSFE shows improvements of 0.10%/0.09%/0.65%/0.83% and 0.24%/0.61%/0.22%/1.08% in the total, head, medium, and tail categories, respectively. The significant improvements in the tail categories further demonstrate that the proposed progressive attention enhancement mechanism greatly enhances the model’s ability to recognize tail classes. This also indirectly validates our hypothesis that the primary impact of long-tailed multi-label data on deep neural networks is insufficient confidence in tail categories.

4.4. Ablation Analysis

This section focuses on an ablation study on the core modules of CSSFE. To this end, we first construct three new sub-models. In detail, Net-0 consists of the backbone and a normal convolutional layer, corresponding to the basic architecture described in Section 3.2. It can be regarded as the baseline. Net-1 adds text clues to the baseline, which can be deemed as CSSFE without the progressive attention enhancement module. Net-2 further incorporates the progressive attention module into Net-1. However, it only ad-

justs global attention through semantic association without introducing the binary mask (see Equation 2) for selective regulation. Apart from the three new sub-models mentioned above, the fully implemented CSSFE is recorded as Net-3 here for the ablation study. Then, we apply them to LT-VOC and LT-COCO and count the classification results, which are summarized in Figure 4.

Upon the observation, we can first confirm that all elements in CSSFE contribute positively. The performance of total categories of Net-0 to Net-3 shows a stepwise improvement across the two datasets. Second, the textual information embedded in the pre-trained language models greatly benefits the long-tailed multi-label image classification task. This is evident in the significant performance improvements of Net-1 compared to Net-0. Third, by comparing Net-2 with Net-1, we find that adjusting the model’s attention to different categories based on semantic relationships is feasible. The overall performance is enhanced, however, this increase comes at the cost of reducing performance on head categories (LT-VOC) or both head and middle categories (LT-COCO) in exchange for improved performance on tail categories. The primary reason behind this is that the number of tail categories is limited, and their semantic relationships often reduce the model’s confidence in other categories, thus affecting the model’s decisions. This highlights the necessity of our unidirectional attention enhancement from head to medium to tail. As a result, Net-3 demonstrates apparent improvements in performance across overall, head, medium, and tail categories.

Due to space limitations, the effects of different feature extraction backbones and label embeddings generated by various pre-trained large-scale language models are studied in the Supplementary Material.

4.5. Further Study

In this paper, we process feature learning analysis of deep neural networks and conclude that when confronted with long-tailed multi-label data, models retain basic localization capabilities but struggle to produce high-confidence scores for tail categories. Motivated by this finding, we propose a progressive attention enhancement method that improves overall classification performance by adjusting the class activation maps for specific categories. To demonstrate the broad applicability of our conclusion and core methodology, we embed our progressive attention enhancement mechanism into the top 3 long-tailed multi-label image classification methods mentioned in Section 4.3 (i.e., Bilateral-TFS, CAE-Net, and CPRFL) and examine if the behavior of new models (marked by †) would be strengthened or not. Note that the new models combine the original feature extractors and classifiers with the progressive attention enhancement mechanism.

The comparison results between new models and their

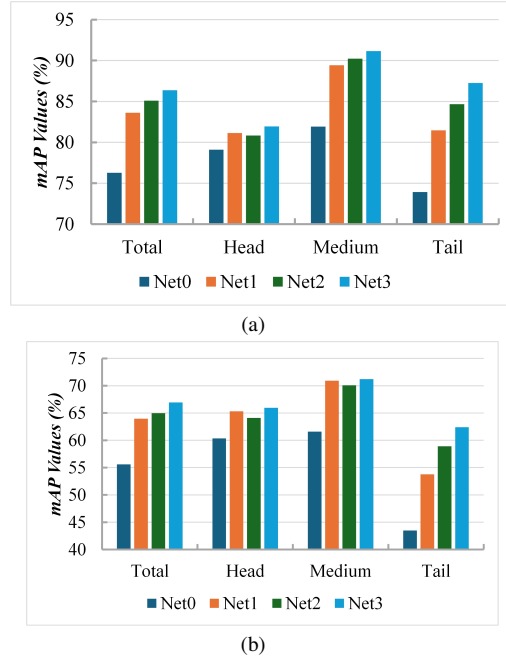


Figure 4. Ablation analysis of the impact of key modules on our proposed CSSFE. (a) on LT-VOC.(b) on LT-COCO.

Table 2. The mAP (%) performance of the top-3 long-tail multi-label image classification methods mentioned in Section 4.3 and three new models (marked by †) using their feature extractors and loss functions combined with our progressive attention enhancement mechanism.

Dataset	LT-VOC			
Methods	total	head	medium	Tail
Bilateral-TFS	81.58	75.88	84.11	83.95
Bilateral-TFS†	85.66	81.04	89.32	86.26
CAE-Net	81.61	74.00	85.35	85.28
CAE-Net†	85.83	81.16	90.41	86.14
CPRFL	86.28	81.84	90.51	86.43
CPRFL†	86.32	81.88	90.78	86.96

Dataset	LT-COCO			
Methods	total	head	medium	Tail
Bilateral-TFS	56.38	55.93	58.26	54.29
Bilateral-TFS†	65.79	65.32	70.24	60.58
CAE-Net	57.64	52.37	61.18	57.63
CAE-Net†	66.18	65.26	70.38	61.26
CPRFL	66.69	66.35	70.99	61.33
CPRFL†	66.82	66.41	71.14	61.48

original versions counted on LT-VOC and LT-COCO are summarized in Table 2. Delightedly, the new models demonstrate consistent improvements in all cases, regardless of the used backbones (including convolutional neural networks and Transformers) and loss functions (such as

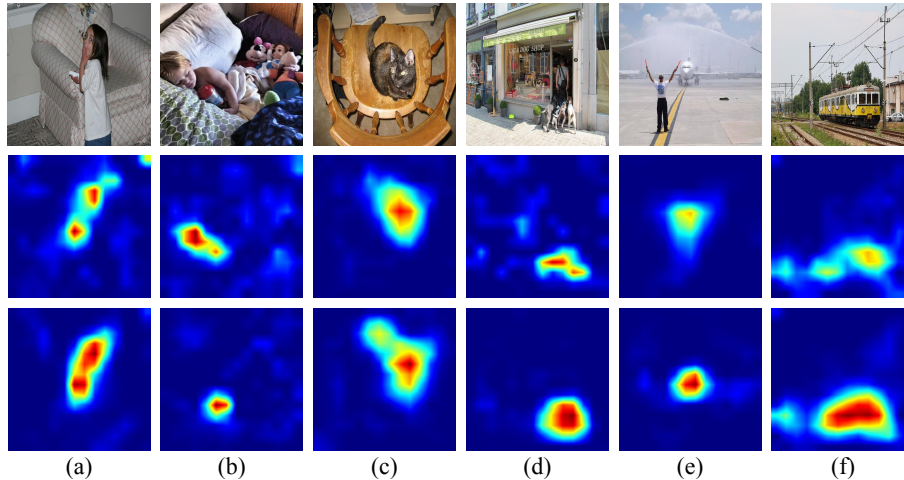


Figure 5. Images in the first row are randomly selected from LT-COCO, the second row is the class activation maps generated by CSSFE without the progressive attention enhancement module, and the third row is the class activation maps generated by the complete CSSFE. (a) Head category: person. (b) Head category: cup. (c) Medium category: cat. (d) Medium category: dog. (e) Tail category: airplane. (f) Tail category: train.

ASL and DB losses). For instance, on the LT-VOC dataset, Bilateral-TFS[†] shows an increase in the overall score from 81.58% to 85.66%, with similar improvements observed in the head, medium, and tail categories. These results indicate that our progressive attention enhancement mechanism successfully re-calibrates the class activation maps, allowing the model to assign higher confidence scores to tail categories while maintaining or even improving the performance on the head and medium categories. It is worth emphasizing that the progressive attention enhancement mechanism is based on the conclusion that although the deep neural network is less sensitive to the features of the tail category, its inherent localization ability can still be effectively activated (see Section 3.2). Therefore, the above-mentioned architecture-independent improvement effect not only verifies the effectiveness of the proposed progressive attention enhancement module but also proves the generalization of the conclusions we analyzed.

4.6. Visualization and Analysis

To better understand how our approach enhances the deep neural network’s ability to learn features from long-tailed multi-label data, we trained two models. One is a partial CSSFE, in which the progressive attention enhancement module is removed. The other one is the complete CSSFE. Then, we visualize their class activation maps across different categories to assess their sensitivity. As shown in Figure 5, the complete CSSFE (the third row) can comprehensively extract features for all classes—particularly for the tail categories—demonstrating a significant improvement in attention compared to the partial CSSFE (the second row). Moreover, the complete CSSFE filters redundant informa-

tion and achieves precise localization. For example, in Figure 5a, the upper right corner of the attention map in the second row has an incorrect activation, which is diluted in the third row. In addition, the attention maps in the third row have significantly fewer spots in the irrelevant area than those in the second row. Overall, our CSSFE significantly improves the feature extraction capabilities of deep neural networks across various classes, yielding notable advantages in long-tailed multi-label image classification tasks.

5. Conclusion

In this paper, we diverge from previous research by investigating the impact of long-tailed distribution on multi-label classification models from the feature learning perspective. Our analysis reveals that deep neural networks can localize class-related regions in long-tailed multi-label image data but lose sensitivity to low-frequency categories. Motivated by this, we propose CSSFE. It first uses the feature extraction backbone to learn the attention area related to the label. Then, it selectively corrects the weights related to the low-frequency categories based on the semantic connection. Finally, the image features and semantic information are combined to complete the classification task. Superior experimental results demonstrate the general applicability of our findings and the proposed method.

6. Acknowledge

This work was supported in part by the National Natural Science Foundation of China under Grant 62171332 and Grant 62276197, in part by the Key Research and Development Program of Shaanxi under Grant 2024GX-YBXM-

125, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2024JC-YBMS-472, in part by Shaanxi Provincial Department of Education Key Scientific Research Project 20JT022, and in part by the Fundamental Research Funds for the Central Universities under Grant YJSJ25004.

References

- [1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arachis, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 6
- [2] Jiayao Chen and Shaoyuan Li. Class-aware learning for imbalanced multi-label classification. In *2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCSIT)*, pages 903–907. IEEE, 2023. 6
- [3] Zhao-Min Chen, Quan Cui, Xiaoqin Zhang, Ruoxi Deng, Chaoqun Xia, and Shijian Lu. Towards gradient equalization and feature diversification for long-tailed multi-label image recognition. *IEEE Transactions on Multimedia*, 2025. 2
- [4] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 6
- [5] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 6
- [6] Ruiqi Du, Xu Tang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Semantic-assisted feature integration network for multi-label remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 1
- [7] Ruiqi Du, Xu Tang, Jingjing Ma, Xiangrong Zhang, and Licheng Jiao. Mlmamba: A mamba-based efficient network for multi-label remote sensing scene classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 1
- [8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5
- [9] Yu Fu, Liuyu Xiang, Yumna Zahid, Guiguang Ding, Tao Mei, Qiang Shen, and Jungong Han. Long-tailed visual recognition with deep models: A methodological survey and evaluation. *Neurocomputing*, 509:290–309, 2022. 2
- [10] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021. 2, 6
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [12] Shu Hu, Xin Wang, and Siwei Lyu. Rank-based decomposable losses in machine learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 13599–13620, 2023. 3
- [13] Kuan-Chih Huang, Ming-Hsuan Yang, and Yi-Hsuan Tsai. Delving into motion-aware matching for monocular 3d object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6909–6918, 2023. 3
- [14] Sontje Ihler, Felix Kuhnke, Timo Kuhlitz, and Thomas Seel. Distribution-aware multi-label fixmatch for semi-supervised learning on chexpert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2295–2304, 2024. 1
- [15] Jaehyup Jeong, Bosoung Jeoun, Yeonju Park, and Bohyung Han. An optimized ensemble framework for multi-label classification on long-tailed chest x-ray data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2739–2746, 2023. 1
- [16] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14156–14165, 2022. 2
- [17] Youngwook Kim, Jae Myung Kim, Jieun Jeong, Cordelia Schmid, Zeynep Akata, and Jungwoo Lee. Bridging the gap between model explanations in partially annotated multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3417, 2023. 3
- [18] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6979, 2022. 2
- [19] Jin-Ha Lim, Myeong-Seok Oh, and Seong-Whan Lee. Enhancing the discriminative ability for multi-label classification by handling data imbalance. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 799–804. IEEE, 2023. 6
- [20] Dekun Lin. Probability guided loss for long-tailed multi-label image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1577–1585, 2023. 2, 6
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3
- [22] Chengliang Liu, Jinlong Jia, Jie Wen, Yabo Liu, Xiaoling Luo, Chao Huang, and Yong Xu. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13864–13872, 2024. 1
- [23] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546, 2019. 6

- [24] Trong-Hieu Nguyen-Mau, Tuan-Luc Huynh, Thanh-Danh Le, Hai-Dang Nguyen, and Minh-Triet Tran. Advanced augmentation and ensemble approaches for classifying long-tailed multi-label chest x-rays. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2729–2738, 2023. 1
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 4, 5
- [26] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021. 6
- [27] Usha Ruby, Vamsidhar Yendapalli, et al. Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng.*, 9(10), 2020. 3
- [28] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21592–21601, 2024. 1
- [29] Pengpeng Song, Anyi Ju, Wenbin Xu, and Fei Guo. Adaptively weighted copy-decoupling resampling strategy for long-tailed multi-label classification. In *2023 IEEE 6th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pages 437–442. IEEE, 2023. 6
- [30] Xin Tan, Ce Zhao, Chengliang Liu, Jie Wen, and Zhanyan Tang. A two-stage information extraction network for incomplete multi-view multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15249–15257, 2024. 1
- [31] Xu Tang, Dabiao Huang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Prior-experience-based vision-language model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 3
- [32] Adane Nega Tarekegn, Krzysztof Michalak, Giuseppe Costa, Fulvio Ricceri, and Mario Giacobini. Predicting multiple outcomes associated with frailty based on imbalanced multi-label classification. *Journal of Healthcare Informatics Research*, pages 1–25, 2024. 1
- [33] Duc-Quang Vu, Trang TT Phung, Jia-Ching Wang, and Son T Mai. Lcsl: Long-tailed classification via self-labeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2
- [34] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9695–9704, 2021. 2
- [35] Yijing Wang, Xu Tang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Cross-modal remote sensing image-text retrieval via context and uncertainty-aware prompt. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 3
- [36] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer, 2020. 2, 3, 5, 6
- [37] Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. Does head label help for long-tailed multi-label text classification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 14103–14111, 2021. 1
- [38] Yosuke Yamagishi and Shohei Hanaoka. Effect of stage training for long-tailed multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2721–2728, 2023. 1
- [39] Jiexuan Yan, Sheng Huang, NanKun Mu, Luwen Huangfu, and Bo Liu. Category-prompt refined feature learning for long-tailed multi-label image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2146–2155, 2024. 2, 6
- [40] Lu Yang, He Jiang, Qing Song, and Jun Guo. A survey on long-tailed visual recognition. *International Journal of Computer Vision*, 130(7):1837–1872, 2022. 2
- [41] Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, and Yueting Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1423–1432, 2023. 2, 6
- [42] Yuhang Zhang, Chengrui Wang, Xu Ling, and Weihong Deng. Learn from all: Erasing attention consistency for noisy label facial expression recognition. In *European Conference on Computer Vision*, pages 418–434. Springer, 2022. 3
- [43] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023. 1, 2
- [44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2, 3
- [45] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 6
- [46] Xuelin Zhu, Jian Liu, Weijia Liu, Jiawei Ge, Bo Liu, and Jixun Cao. Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1473–1482, 2023. 1