future internet

MDPI

Article

Class-Level Feature Disentanglement for Multi-Label Image Classification

Yingduo Tong [1], Zhenyu Lu [2], Yize Dong [1] and Yonggang Lu [1,*]

[1] School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China; tongyd2023@lzu.edu.cn (Y.T.); dyize2023@lzu.edu.cn (Y.D.)
[2] School of Educational Technology, NorthWest Normal University, Lanzhou 730070, China; lizziy@nwnu.edu.cn
* Correspondence: ylu@lzu.edu.cn

Abstract

Generally, the interpretability of deep neural networks is categorized into a priori and a posteriori interpretability. A priori interpretability involves improving model transparency through deliberate design prior to training. Feature disentanglement is a method for achieving a priori interpretability. Existing disentanglement methods mostly focus on semantic features, such as intrinsic and shared features. These methods distinguish between the background and the main subject, but overlook class-level features in images. To address this, we take a further step by advancing feature disentanglement to the class level. For multi-label image classification tasks, we propose a class-level feature disentanglement method. Specifically, we introduce a multi-head classifier within the feature extraction layer of the backbone network to disentangle features. Each head in this classifier corresponds to a specific class and generates independent predictions, thereby guiding the model to better leverage the intrinsic features of each class and improving multi-label classification precision. Experiments demonstrate that our method significantly enhances performance metrics across various benchmarks while simultaneously achieving a priori interpretability.

Keywords: priori interpretability; feature disentanglement; multi-head classifier; image classification

check for updates

Academic Editor: Paolo Bellavista

Received: 18 September 2025
Revised: 16 October 2025
Accepted: 20 October 2025
Published: 23 October 2025

Citation: Tong, Y.; Lu, Z.; Dong, Y.; Lu, Y. Class-Level Feature Disentanglement for Multi-Label Image Classification. *Future Internet* **2025**, *17*, 486. https://doi.org/10.3390/fi17110486

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, deep learning has made remarkable progress in image classification tasks. However, most existing methods still rely on end-to-end black-box models, lacking interpretability in their feature representations [1]. Traditional classifiers typically learn global feature representations, making it difficult to distinguish between intrinsic features and shared features. This limitation restricts the generalizability and interpretability of the model in complex scenarios. For example, in medical image analysis, the intrinsic morphological features of lesions may be obscured by noise or differences in imaging equipment [2]; in fine-grained classification tasks, subtle interclass differences are often masked by shared features [3].

Existing research has attempted to enhance model interpretability through feature disentanglement; generative models like VAEs and GANs separate semantically distinct features via latent space decomposition [4]. However, these methods usually depend on unsupervised or weakly supervised learning, making it challenging to precisely control the granularity of feature separation in classification tasks. Multi-head classifiers with their parallel learning capability offer a novel approach to feature disentanglement; by designing

task-specific heads, they can explicitly model different feature subspaces [5]. However, current approaches have not fully explored the explicit separation mechanism between intrinsic features and shared features, resulting in a lack of transparency in the classification decision-making process [6].

Existing disentanglement methods primarily focus on separating intrinsic features and shared features. Intrinsic features mostly represent characteristics of the main subject and serve as the basis for determining the object's class, while shared features often correspond to background elements. These shared features are frequently mixed with those of multiple classes that share the same background, which can negatively impact the accuracy of classification tasks. By separating intrinsic features from shared features, the reliability of classification tasks can be improved [7].

Our method adheres to the principle of a priori interpretability and directly targets class-level features, proposing a new direction for feature disentanglement. The distinguished features are no longer limited to the intrinsic and shared features of individual classes, but directly differentiate the overall features between classes. By emphasizing the importance of class information, the designed multi-head classifier focuses on class-level feature disentanglement. Subsequent experiments also demonstrate that class-level feature disentanglement improves the accuracy of image multi-label classification.

To address these challenges, this study proposes a priori interpretable feature disentanglement architecture based on a multi-head classifier. Currently, most research on interpretability focuses on a posteriori interpretability, methods like latent layer analysis, surrogate models, and sensitivity analysis. In contrast, priori interpretability enhances model trustworthiness, reduces interpretation bias, and improves debugging efficiency. Our work takes a class-level feature disentanglement approach for image multi-label classification tasks and integrates priori interpretability. Our contributions are mainly organized as follows:

1. We propose an Explicit Decoupling Mechanism in the Feature Extraction Layer. Specifically, a multi-head classifier is designed to separate intrinsic features from shared features during the feature extraction stage. This design endows the model with a priori interpretability.

2. We propose a Joint Optimization Strategy. Each classification loss corresponds with a classification head, and is optimized independently. This strategy helps to improve the performance of feature disentanglement.

3. We propose an Interpretability-driven Supervision Strategy. Heatmaps are employed to monitor the focal regions of intrinsic features targeted by each classification head, achieving interpretable feature disentanglement.

4. The proposed multi-head classifier allows flexible adjustment of the number of classification heads and their connection strategies, such as hierarchical feature division or channel-wise feature splitting, to accommodate diverse multi-label classification scenarios, showing strong scalability

2. Related Work

2.1. Interpretability

In the field of interpretability research for deep neural networks, scholars have extensively explored methods to uncover the logic and rationale behind model decisions, aiming to enhance model transparency, trustworthiness, and real-world applicability. Early studies primarily focused on visualizing key regions of interest through feature visualization techniques to provide intuitive explanations. For instance, Zeiler and Fergus proposed Deconvolutional Networks, which reverse-map convolutional layer feature maps back to the input space, visually revealing the hierarchical visual patterns captured by the net-

work [8]. Similarly, Simonyan et al. generated Saliency Maps by computing gradients of output class scores with respect to input images, highlighting pixel regions most influential to classification decisions [9]. While these methods offer intuitive visual interpretations, they often lack in-depth analysis of internal mechanisms and suffer from instability due to gradient-based noise sensitivity, compromising the reliability of explanations.

As research progressed, attention mechanism-based interpretability methods gained traction. These approaches explicitly represent the model's focus on different input components during processing, providing structured explanations for decision-making. For example, self-attention weights in Transformer architectures directly reflect the model's information preference across temporal or spatial dimensions. Vaswani et al. visualized attention matrices to demonstrate how the model dynamically associates key elements in input sequences [10]. In computer vision, Dosovitskiy et al.'s Vision Transformer(ViT) leverages attention weights to explain the contribution of image patches to final classifications [11]. However, the interpretability of attention weights remains debated—studies suggest that attention distributions may correlate with but not strictly causally determine model decisions, often reflecting associations rather than definitive decision logic [12].

Beyond visualization and attention mechanisms, feature attribution methods further quantify the contribution of input features to model outputs, offering numerical explanations. Representative approaches include LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). LIME trains a local interpretable proxy model within the input neighborhood to approximate the decision boundary of the complex original model, identifying critical input features for specific predictions [13]. SHAP, grounded in cooperative game theory's Shapley values, allocates marginal contributions to each feature from a multi-feature collaboration perspective, with theoretically robust mathematical foundations [14]. Although these methods improve explanatory rigor, they often entail high computational costs and struggle to directly link attribution results to internal structures, such as convolutional kernel weights or intermediate feature maps, due to the non-linear complexities of deep neural networks.

For architecture-specific interpretability, researchers have investigated intermediate feature analysis to reveal decision bases. Zhou et al.'s Class Activation Mapping technique, for example, reverse-weights convolutional feature maps using fully connected layer weights after the global average pooling layer, generating heatmaps that localize discriminative regions for target classes. Subsequent Grad-CAM refines weight allocation via gradient information, enabling fine-grained visual explanations across broader CNN architectures. These methods directly bridge intermediate feature representations and final classification decisions, providing key insights into how CNNs leverage local visual patterns.

Additionally, research on global model behavior interpretability focuses on uncovering overarching decision logic. Examples include analyzing decision boundaries, clustering characteristics of feature spaces, such as t-SNE/UMAP visualizations of high-dimensional feature distributions, or constructing proxy models to approximate the input–output relationships of complex networks. While these macroscopic approaches supplement understanding of model behavior, they often sacrifice direct associations with internal parameters.

Collectively, interpretability research for deep neural networks has evolved into a multi-dimensional, hierarchical technical framework [15]. Yet core challenges persist: How to balance intuitive accessibility with rigorous explanatory depth? How to reconcile computational efficiency with explanatory granularity? And how to design task-specific and architecture-adapted interpretability methods? Current trends are shifting toward co-optimization of interpretability and model performance, such as integrating priori interpretable modules or embedding interpretability constraints during training, like mini-

mizing adversarial perturbations or enforcing feature consistency regularization. These advancements are driving deep learning models from "black boxes" toward "gray boxes" and even "white boxes," providing critical technical support for establishing user trust and meeting compliance requirements in high-stakes applications, for instance, in medical diagnostics [16] and autonomous driving.

2.2. Feature Disentanglement

As a pivotal technique for enhancing model generalization, interpretability, and controllability, feature disentanglement in deep neural networks has emerged as a research frontier in recent years. Traditional feature engineering methods, relying on manually designed rules or domain-specific prior knowledge to separate mixed features, such as Principal Component Analysis (PCA) extracting orthogonal principal components via linear transformations, struggle to adapt to complex non-linear data distributions like multisemantic coupled features in images [17]. With the rise of deep learning, researchers have shifted toward end-to-end learning for automated feature disentanglement, aiming to map input data into a low-dimensional semantic space where features across different dimensions correspond exclusively to singular semantic attributes. This approach improves model robustness and controllability for specific tasks.

Early disentanglement methods predominantly leveraged the latent space structures of generative adversarial networks (GANs) or variational autoencoders (VAEs). For instance, InfoGAN maximized the mutual information between latent variables and generated data, disentangling the VAE latent space into continuous semantic factors and irrelevant noise under unsupervised conditions [18]. However, its reliance on generative model stability and the predefined prior assumptions of latent space structures limited the controllability of disentangled features. Similarly, β-VAE introduced a KL divergence weight coefficient β to regulate the independence of latent variables, achieving partial separation of semantic attributes in image reconstruction tasks [19]. Yet, it faced challenges in precisely controlling the coupling strength of multidimensional features. While these methods laid the theoretical foundation for latent space disentanglement, practical applications often encountered issues such as semantically ambiguous disentangled features and insufficient cross-task generalization.

To address more complex visual tasks, researchers proposed explicit disentanglement strategies based on discriminative models. One representative approach utilized contrastive learning to separate shared and specific features. By constructing multiview representations of the same sample through data augmentation, models were constrained to map shared semantics into aligned feature spaces while isolating view-dependent features into orthogonal subspaces. Frameworks like SimCLR and MoCo, for example, employed contrastive losses to pull positive sample pairs closer and push negative pairs apart, implicitly disentangling category-relevant features from irrelevant ones in unlabeled data [20]. However, their disentanglement process depended on large-scale data augmentation and lacked direct control over specific semantic factors. Another line of work adopted multi-branch network architectures to explicitly separate distinct semantic dimensions. For fine-grained image classification, dual-stream networks were developed—one branch extracted global category features, while the other focused on local discriminative regions. By fusing features via concatenation or attention-weighted mechanisms, these methods effectively disentangled mixed features into global and local semantic components, thereby improving classification accuracy.

Recent advances have emphasized lightweight and task-driven feature disentanglement to reduce computational complexity and enhance efficiency. Techniques such as grouped convolution, sparse activation, and low-rank decomposition have been employed to constrain feature representations. For example, in multi-head classifiers, independent lightweight convolutional heads were designed for each category, limiting parameter counts to force each head to focus solely on locally relevant features strongly associated with its target class [21]. This indirectly achieved disentanglement from input features to category-discriminative factors. Similarly, dynamic convolution generated adaptive convolutional kernel weights for different input samples, dynamically adjusting feature extraction paths during inference to activate only task-relevant semantic components and mitigate interference from redundant features. Additionally, causality-informed disentanglement methods constrained models to learn latent features invariant to environmental changes, disentangling source-domain deviations from target-domain-specific attributes in cross-domain generalization tasks [22]. This significantly improved model robustness in unseen scenarios.

These disentanglement techniques demonstrate unique advantages across applications. In medical imaging analysis, they separate lesion features from normal tissue backgrounds, aiding clinicians in precise lesion localization; in facial recognition, disentangling identity features from environmental factors enhances adaptability to complex scenes; in autonomous driving, isolating road condition features from weather interferences strengthens perception system reliability [23].

Collectively, research on feature disentanglement in deep neural networks has evolved from early unsupervised latent space explorations to explicit semantic control and task-driven optimization. Its core objective remains the structured separation of independent semantic components within mixed features, providing models with more transparent and controllable feature representations. This progression drives deep learning toward interpretable AI, transitioning from "black boxes" to systems with explainable and trustworthy foundations.

3. Method

3.1. Overall Model Architecture

The multi-head classifier proposed in this study introduces additional classification head modules on the basis of traditional backbone networks to improve the Top-k accuracy. Its core concept is to enable more effective utilization of information from different feature layers through hierarchical feature disentanglement and joint optimization. This is shown in Figure 1. The overall model architecture consists of three main components:

1. Backbone Network. In this study, VGG-19, SE-ResNet50, and Swin Transformer are selected as the backbone networks to extract multi-level features.

2. Multi-Head Classifier Module. The multi-head classifier is inserted after the feature layers of the backbone network, replacing the original fully connected layer classifier. The features are divided into multiple subsets and processed by independent classification heads to achieve hierarchical feature disentanglement.

3. Prediction Fusion Module. The outputs of each classification head are integrated, where the one-dimensional tensor output by each head represents the prediction result for the corresponding class label, and the results from all heads are fused to form the final prediction.

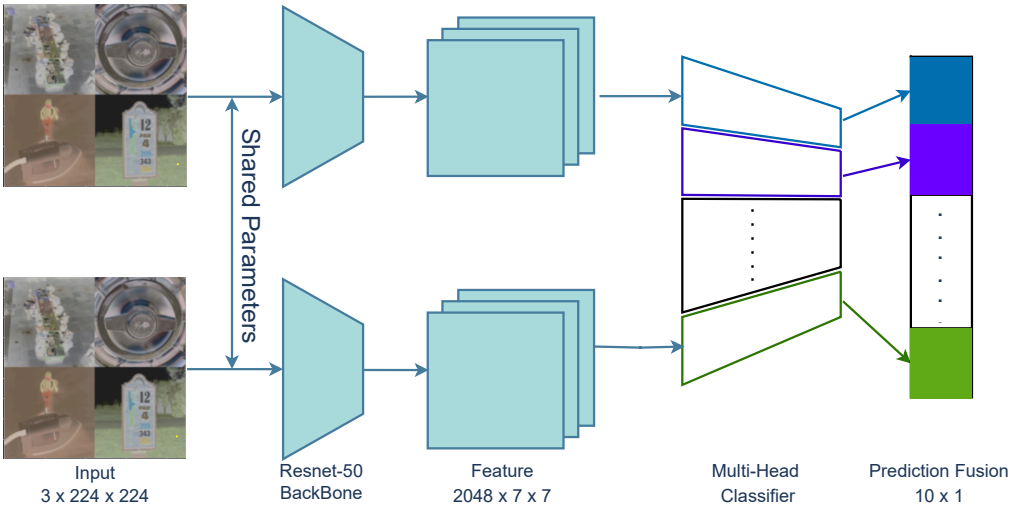


Figure 1. The input image is first processed by the ResNet backbone network to obtain the feature maps. After the feature maps, a multi-head classifier module is inserted. This module contains multiple parallel classification heads. Each classification head independently calculates and outputs the classification probability, and then the fusion module generates the final prediction.

3.2. Design of Multi-Head Classifier

The multi-head classifier proposed in this study performs class-level partitioning on the features extracted by the backbone network. Each classification head is responsible for its corresponding class labels and learns features specific to that category. Each head maintains independent parameters when the multi-head classifier learns class labels by minimizing the binary cross-entropy that measures the difference between integrated predicted probabilities and class labels, thereby achieving feature disentanglement. This is shown in Figure 2.

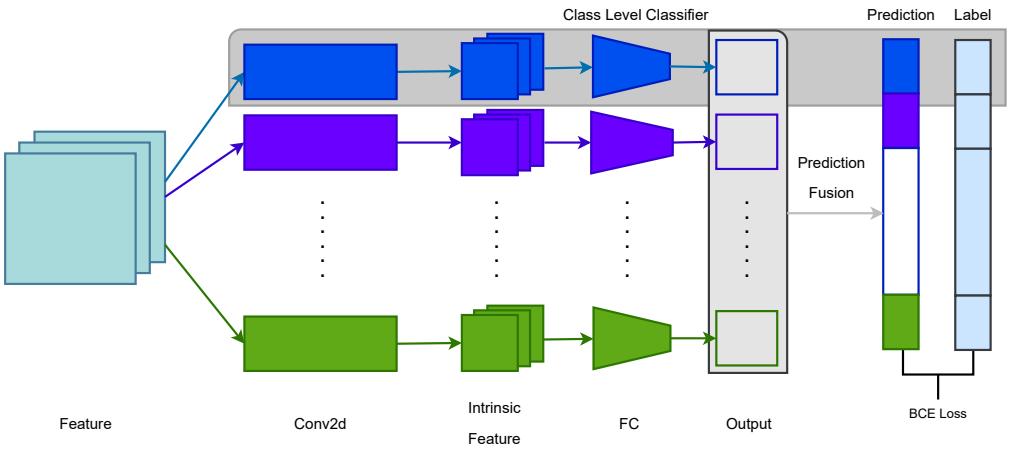


Figure 2. Multi-Head Classifier. The design concept embodies priori interpretability, where each classification head is dedicated to its corresponding class label, enabling class-level feature disentanglement.

Each head shares identical structure but independent parameters. Each classification head comprises lightweight convolutional layers to assist in feature extraction, along with an average pooling layer and a fully connected layer to generate prediction results. Certainly, the convolutional layers are optional. The fully connected layers, which are the core of the classifier, can also demonstrate the effectiveness of the multi-head classifier.

3.3. Overview of the Proposed Framework

To clearly delineate our contributions from existing works, this subsection provides a high-level overview of the proposed framework. The core innovation of this study lies in a novel multi-head classifier architecture designed for a priori interpretability and class-level feature disentanglement. Figure 3 illustrates the overall pipeline of our proposed method.

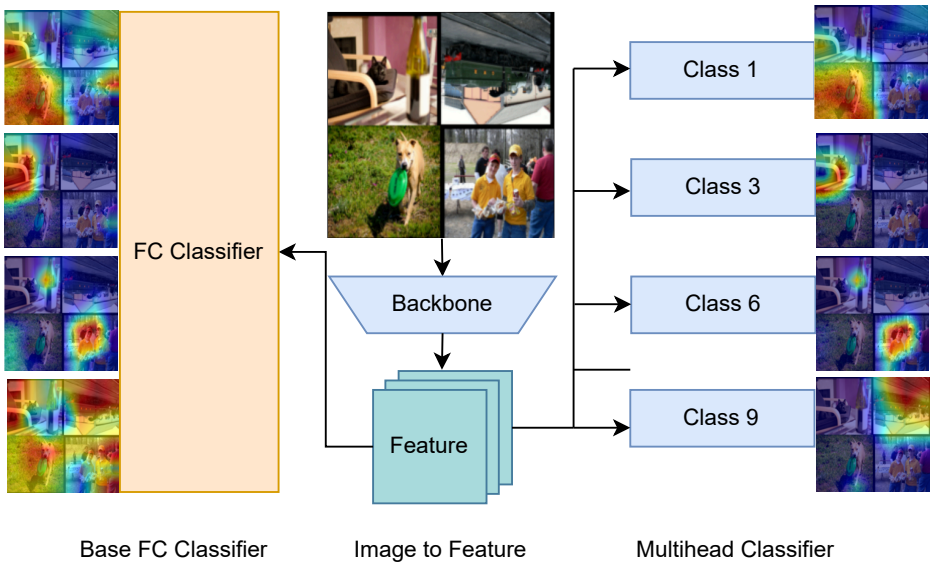


Figure 3. End-to-end workflow from input image to heatmap of interpretable multi-label predictions.

3.3.1. The Gap in Existing Methods

While current deep learning models for multi-label classification have shown remarkable performance, they often function as black boxes, lacking transparency in how predictions for multiple labels are derived. Specifically, standard classifiers learn entangled feature representations, making it difficult to trace which part of the model is responsible for which class.

3.3.2. Our Core Idea

To bridge this gap, we propose a divide-and-conquer strategy at the classifier level. The fundamental premise is to assign a dedicated, independent classification head to each class label. This design explicitly forces the model to learn and maintain separate feature representations for each class, thereby achieving class-level feature disentanglement by construction.

3.3.3. Our Specific Creations

The key component we novelly introduced or significantly adapted in this work is the Multi-Head Classifier Block. This is our primary architectural contribution. Unlike a standard single-head classifier, our block comprises N independent parallel classification heads where N is the number of classes, each with its own set of parameters.

4. Experiment

4.1. Datasets

To investigate the principle and effectiveness of feature disentanglement in the multi-head classifier, this study introduces a multi-label dataset constructed by splicing images from a multi-class dataset. The images in the multi-class dataset contain only a single class and its corresponding subject. We selected single-label images from a total of 10 categories

and spliced four images with different class labels into a 2 × 2 multi-label dataset. This is shown in Figure 4.



Figure 4. Combined datasets. This is a spliced image that mimics multi-label classification images. It makes the features relatively concentrated, excludes background interference, and meanwhile incorporates more coupled class-level features.

Leveraging the characteristics of the combined dataset, we employed Grad-CAM to observe the learning effect of the multi-head classifier on class-level feature disentanglement. The images in the combined dataset have relatively concentrated features, less background interference, and more coupled features, which can better validate the effectiveness of the multi-head classifier.

Before training, the feature heatmap of a certain class in the image marked by Grad-CAM is shown in Figure 5, which can point to the class but is relatively scattered. After training, the feature heatmap of the corresponding class becomes more accurate and has more precise coverage. This is shown in Figure 6. This proves that the multi-head classifier has learned the ability of class-level feature disentanglement on the combined dataset we specifically proposed, which also lays the foundation for subsequent experiments on native multi-label datasets.

After achieving preliminary results on the combined dataset, we then selected two multi-label image classification datasets, namely, Pascal VOC and Open Images, to verify the robustness of feature disentanglement in the multi-head classifier for multi-label image classification tasks. The Pascal VOC dataset contains 20 classes, while OpenImages has 1000 classes. To facilitate experimental execution and result comparison, we performed a screening process for each dataset. Specifically, we selected a total of 10 classes from each dataset, with each image containing 2–7 classes, striving to maintain relatively uniform

distribution across all selected classes. Ultimately, we extracted 10,000 multi-label classification images from each dataset. However, this inevitably resulted in class imbalance. Therefore, for subsequent evaluation metric calculations, we adopted macro-averaging for all metrics.

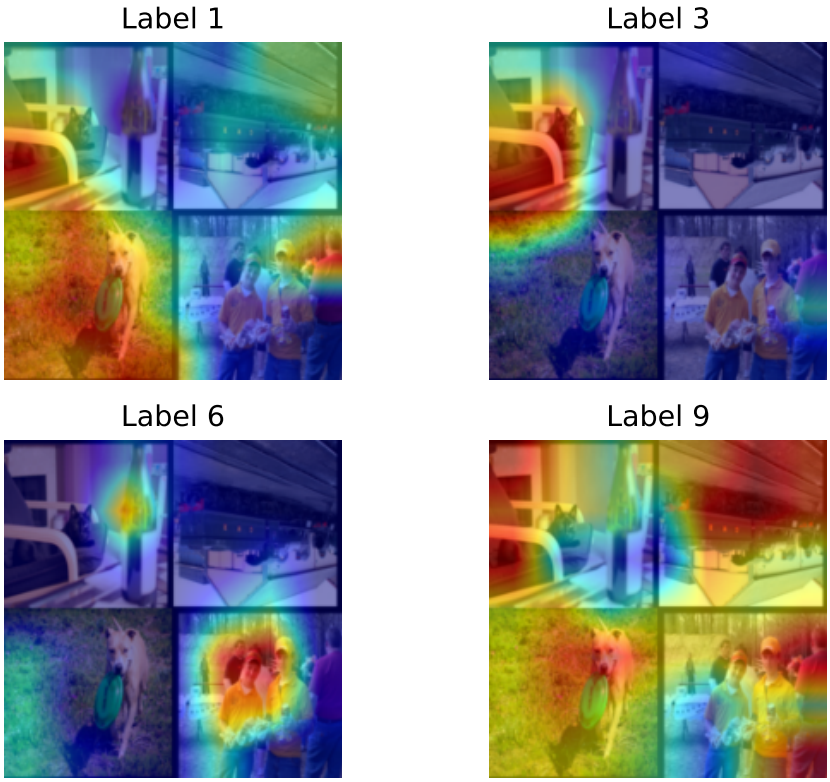


Figure 5. Grad-CAM feature heatmap before training. The pointing is ambiguous and the coverage is imprecise.

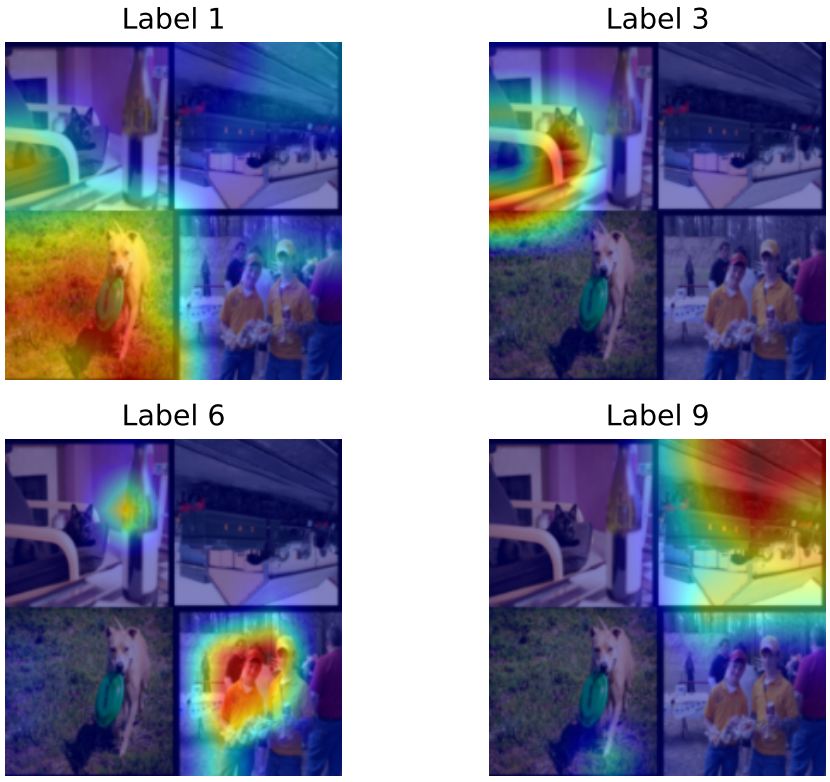


Figure 6. Grad-CAM feature heatmap after training. The pointing is well defined and the coverage is accurate.

4.2. Training Setup

For this study, Se-ResNet50, VGG19, and Swin Transformer were selected as the backbone networks. All parameters of the backbone networks were frozen. The initial learning rate was set to 0.01, with a cosine annealing strategy employed to decay the learning rate, and the SGD optimizer was utilized. For native datasets, the model was trained for 1000 epochs; for the combined dataset, training was conducted for 300 epochs. The batch size for all datasets was set to 128. Data augmentation is basic. Firstly, the image is resized to 224 by 224 pixels to adapt the models; secondly, the image is randomly cropped into 224 by 224 pixels and padded 8 pixels with zero values. Lastly, the image is randomly flipped horizontally.

4.3. Comparison Results

Our baseline is a simple fully connected layer. We designed two distinct multi-head classifier models for comparison with the baseline: the first is a multi-head classifier consisting solely of fully connected layers, designated as MultiFC; the second is a multi-head classifier with convolutional layers preceding the fully connected layers, designated as MultiConv. For control, we also implemented a single-head classifier with convolutional layers before the fully connected layers, designated as ConvFC. In total, four model groups were compared in four datasets: two combined datasets and two native datasets.

We implemented common evaluation metrics for multi-label image classification, including Accuracy, Precision, F1-score, Recall, and Hamming loss (HL). For the combined dataset with balanced classes, micro-averaging was adopted; for the native datasets with imbalanced classes, macro-averaging was employed. Following standard multi-label classification practices, we utilized the top-k method to identify correctly predicted items. For native datasets, results from top-1 to top-5 were recorded, with representative top-1 and top-5 outcomes presented. For the combined dataset where each image definitively contains four classes, the top-4 method was applied; since Precision, F1-score, and Recall yielded identical values in this context, only Precision metrics are reported.

The results of the combined datasets are shown in Tables 1 and 2. The baseline single-head classifier's metrics, when evaluated on traditional convolutional neural networks ResNet and VGG, demonstrated significant improvements after adopting both multi-head classifier variants, which are MultiFC and MultiConv, confirming that the multi-head classifier effectively enhances feature disentanglement. However, on the Swin Transformer architecture, the MultiFC showed minimal improvement, and the ConvFC even exhibited negative effects. This indicates that standalone convolutional layers alone do not contribute to feature disentanglement gains. Nevertheless, a direct comparison between ConvFC and MultiConv reveals that when a single classification head performs poorly, the multi-head classifier can leverage its advantages in feature disentanglement learning capabilities to achieve superior classification results.

The results of the native datasets are shown in Tables 3 and 4. Since multi-label classification involves multiple classes per image, the Top-1 result is not representative; Top-5 better reflects the performance of the multi-head classifier. Consistent with the combined dataset findings, on ResNet and VGG, MultiFC achieved the best results, with more pronounced improvements from the multi-head classifier. This demonstrates that the multi-head classifier also exhibits feature disentanglement capabilities in native multi-label image classification datasets. Standalone convolutional layers offer no benefit for feature extraction, and while MultiConv partially mitigated the negative effects of ConvFC, this indirectly confirms the feature disentanglement advantages of the multi-head classifier. On the Swin Transformer model, the multi-head classifier generated positive outcomes across all evaluation metrics, validating its robustness and generalizability.

Table 1. Combined Dataset0.

Top-4	Model	Acc (%)	Pre (%)	HL (%)
ResNet	Baseline	59.91	49.88	40.09
	ConvFC	59.86	49.85	40.12
	MultiFC	**63.20**	**54.02**	**36.93**
	MultiConv	63.07	53.85	37.78
VGG	Baseline	58.42	48.22	41.42
	ConvFC	59.91	49.88	40.09
	MultiFC	**60.92**	**51.14**	**39.09**
	MultiConv	60.24	50.30	39.76
SwinT	Baseline	62.24	52.79	37.76
	ConvFC	59.91	49.88	40.09
	MultiFC	**62.27**	**52.83**	**37.73**
	MultiConv	61.65	52.06	38.34

Note: Bold values indicate the best performance on that metric.

Table 2. Combined Dataset1.

Top-4	Model	Acc (%)	Pre (%)	HL (%)
ResNet	Baseline	52.15	40.18	47.85
	ConvFC	52.15	40.18	47.84
	MultiFC	67.02	58.77	**32.98**
	MultiConv	**67.05**	**58.78**	32.99
VGG	Baseline	52.06	40.07	47.94
	ConvFC	52.06	40.07	47.94
	MultiFC	**63.41**	**54.27**	**36.58**
	MultiConv	61.89	52.35	38.12
SwinT	Baseline	70.22	62.77	29.78
	ConvFC	51.96	39.95	48.04
	MultiFC	**70.28**	**62.85**	**29.72**
	MultiConv	69.91	62.39	30.09

Note: Bold values indicate the best performance on that metric.

Table 3. Pascal Voc.

	Top-k	Model	Acc (%)	Pre (%)	F1 (%)	Rec (%)	HL (%)
ResNet	Top-1	Baseline	82.30	7.59	8.64	10.00	17.73
		ConvFC	82.30	7.59	8.64	10.00	17.73
		MultiFC	82.30	7.59	8.64	10.00	17.73
		MultiConv	82.30	7.59	8.64	10.00	17.73
	Top-5	Baseline	62.81	23.14	30.38	53.98	37.19
		ConvFC	61.48	20.25	25.68	50.53	38.52
		MultiFC	**66.77**	31.20	**42.42**	**71.05**	**33.23**
		MultiConv	65.92	**31.46**	41.63	67.73	34.08
VGG	Top-1	Baseline	82.30	7.59	8.64	10.00	17.73
		ConvFC	82.30	7.59	8.64	10.00	17.73
		MultiFC	82.81	27.79	10.34	10.82	17.19
		MultiConv	82.30	7.59	8.64	10.00	17.73
	Top-5	Baseline	63.81	29.89	39.27	63.90	36.19
		ConvFC	61.57	21.48	28.17	51.20	38.43
		MultiFC	**67.01**	**32.04**	**44.02**	**75.70**	**32.99**
		MultiConv	66.34	30.71	41.72	69.11	33.63
SwinT	Top-1	Baseline	82.63	36.9	9.95	10.61	17.37
		ConvFC	82.30	7.60	8.64	10.00	17.7
		MultiFC	82.48	27.67	9.32	10.33	17.55
		MultiConv	82.30	17.61	8.87	10.06	17.73
	Top-5	Baseline	67.64	33.35	45.26	74.85	32.36
		ConvFC	60.79	17.53	23.71	49.85	39.21
		MultiFC	**68.31**	34.29	46.51	**76.87**	**31.69**
		MultiConv	67.98	**35.12**	**46.68**	75.42	32.02

Note: Bold values indicate the best performance on that metric.

Table 4. OpenImages.

	Top-k	Model	Acc (%)	Pre (%)	F1 (%)	Rec (%)	HL (%)
ResNet	Top-1	Baseline	66.94	6.28	7.72	10.00	33.06
		ConvFC	66.94	6.28	7.72	10.00	33.06
		MultiFC	68.22	44.77	11.46	11.43	31.83
		MultiConv	67.47	45.23	9.54	10.63	32.51
	Top-5	Baseline	65.72	25.67	33.83	50.00	34.28
		ConvFC	65.72	25.67	33.83	50.00	34.28
		MultiFC	**70.31**	41.92	**43.14**	**56.77**	**29.73**
		MultiConv	70.23	**43.55**	42.96	56.59	29.77
VGG	Top-1	Baseline	67.77	14.46	10.33	10.98	31.67
		ConvFC	66.97	14.95	7.95	10.06	32.99
		MultiFC	71.49	45.37	19.99	15.43	28.50
		MultiConv	70.36	44.65	16.20	14.00	29.55
	Top-5	Baseline	63.03	24.33	32.51	50.00	36.97
		ConvFC	63.03	24.33	32.51	50.00	36.97
		MultiFC	**73.31**	48.58	**50.52**	**62.47**	**26.70**
		MultiConv	70.11	**52.84**	43.14	57.18	29.89
SwinT	Top-1	Baseline	73.08	57.12	22.94	17.12	26.92
		ConvFC	66.94	6.28	7.72	10.00	33.06
		MultiFC	73.17	47.11	22.86	17.22	26.84
		MultiConv	72.68	46.76	21.40	16.52	27.32
	Top-5	Baseline	74.36	46.97	51.07	63.59	25.67
		ConvFC	65.72	25.67	33.83	50.00	34.28
		MultiFC	**74.32**	**47.57**	**51.14**	**63.50**	**25.65**
		MultiConv	73.14	46.40	48.99	61.33	26.86

Note: Bold values indicate the best performance on that metric.

4.4. Ablation Study

We conducted ablation experiments on the learning rate, batch size, and loss function coefficient. For the learning rate, we compared three values: 0.1, 0.01, and 0.001. For the batch size, we compared three settings: 64, 128, and 256. For the loss function coefficient, we compared three values: 0.5, 0.7, and 1.0. After careful consideration, we chose a learning rate of 0.01, a batch size of 128, and set the loss function coefficient λ to 1 for our experiments. The ablation experiments were performed solely on the best-performing MultiFC architecture and representing metrics such as Accuracy, Precision, and Hamming Loss (HL). As ResNet demonstrated the best performance among the backbones, we selected it for comparative ablation experiments. The results of the learning rate ablation experiment are presented in Table 5. The results of the batch size ablation experiment are presented in Table 6. The results of the learning rate ablation experiment are presented in Table 7.

Table 5. Learning rate ablation.

Dataset	Learning Rate	Model	Acc (%)	Pre (%)	HL (%)
Combined Dataset0	0.1	Baseline	59.23	49.12	39.45
		MultiFC	62.55	53.37	36.28
	0.01	Baseline	59.91	49.88	40.09
		MultiFC	63.20	54.02	36.93
	0.001	Baseline	58.34	48.23	38.76
		MultiFC	61.78	52.45	35.21
Combined Dataset1	0.1	Baseline	51.67	39.71	47.12
		MultiFC	66.23	58.01	32.34
	0.01	Baseline	52.15	40.18	47.85
		MultiFC	67.02	58.77	32.98
	0.001	Baseline	50.76	38.89	46.33
		MultiFC	65.23	57.12	31.45

Table 5. *Cont.*

Dataset	Learning Rate	Model	Acc (%)	Pre (%)	HL (%)
Pascal VOC	0.1	Baseline	62.15	22.67	36.71
		MultiFC	66.45	30.67	32.71
	0.01	Baseline	62.81	23.14	37.19
		MultiFC	66.77	31.20	33.23
	0.001	Baseline	61.12	21.78	35.78
		MultiFC	65.12	29.67	30.67
OpenImages	0.1	Baseline	64.91	25.01	33.67
		MultiFC	67.55	44.01	31.21
	0.01	Baseline	65.72	25.67	34.28
		MultiFC	68.22	44.77	31.83
	0.001	Baseline	63.23	23.12	32.67
		MultiFC	66.67	43.12	29.12

Table 6. Batch size ablation.

Dataset	BatchSize	Model	Acc (%)	Pre (%)	HL (%)
Combined Dataset0	64	Baseline	59.65	49.54	40.21
		MultiFC	63.13	54.11	36.62
	128	Baseline	59.91	49.88	40.09
		MultiFC	63.20	54.02	36.93
	256	Baseline	59.45	49.93	40.13
		MultiFC	62.90	53.89	36.56
Combined Dataset1	64	Baseline	52.24	40.24	47.83
		MultiFC	66.78	59.09	32.52
	128	Baseline	52.15	40.18	47.85
		MultiFC	67.02	58.77	32.98
	256	Baseline	52.31	40.12	47.13
		MultiFC	67.12	58.55	32.58
Pascal VOC	64	Baseline	62.92	23.03	37.34
		MultiFC	66.87	31.52	33.54
	128	Baseline	62.81	23.14	37.19
		MultiFC	66.77	31.20	33.23
	256	Baseline	62.55	23.54	37.12
		MultiFC	66.42	31.09	33.18
OpenImages	64	Baseline	65.21	25.15	34.24
		MultiFC	68.54	44.84	32.02
	128	Baseline	65.72	25.67	34.28
		MultiFC	68.22	44.77	31.83
	256	Baseline	65.24	25.56	34.99
		MultiFC	68.45	44.19	32.32

Table 7. Lambda ablation.

Dataset	Lambda	Model	Acc (%)	Pre (%)	HL (%)
Combined Dataset0	0.5	Baseline	60.05	50.04	40.15
		MultiFC	63.11	53.57	37.34
	0.7	Baseline	59.86	49.76	40.54
		MultiFC	63.05	52.98	37.12
	1.0	Baseline	59.91	49.88	40.09
		MultiFC	63.20	54.02	36.93
Combined Dataset1	0.5	Baseline	52.35	41.38	46.42
		MultiFC	66.23	55.65	33.52
	0.7	Baseline	52.96	41.27	46.02
		MultiFC	66.67	57.24	37.28
	1.0	Baseline	52.15	40.18	47.85
		MultiFC	67.02	58.77	32.98

Table 7. *Cont.*

Dataset	Lambda	Model	Acc (%)	Pre (%)	HL (%)
Pascal VOC	0.5	Baseline	63.53	27.16	36.22
		MultiFC	65.20	31.08	34.44
	0.7	Baseline	62.98	24.07	36.94
		MultiFC	65.41	30.27	34.58
	1.0	Baseline	62.81	23.14	37.19
		MultiFC	66.77	31.20	33.23
OpenImages	0.5	Baseline	66.15	26.18	33.25
		MultiFC	67.02	43.35	32.18
	0.7	Baseline	66.06	25.98	33.94
		MultiFC	68.10	44.34	31.89
	1.0	Baseline	65.72	25.67	34.28
		MultiFC	68.22	44.77	31.83

5. Discussion

5.1. Backbone Network and Multi-Head Classifier

With the rapid advancement in deep learning technology, convolutional neural networks (CNNs) have achieved remarkable success in computer vision tasks. Among these, ResNet effectively alleviates the vanishing gradient problem in deep networks by introducing residual connections, establishing itself as a foundational architecture for numerous vision tasks. However, despite ResNet's outstanding performance in extracting multi-level features, its fully connected (FC) layers exhibit limitations in processing high-dimensional features:

1. Feature Redundancy and Confusion. The deep features of ResNet, after undergoing multiple convolutions and pooling operations, while encompassing rich semantic information, may exhibit redundancy across hierarchical levels or semantic confusion. The fully connected layers typically apply Global Average Pooling (GAP) directly to these features and then rely on a single classification head for final predictions. This approach may fail to fully exploit the complementary nature of features across different layers.

2. Limitations of Single-Path Classification. In traditional ResNet, the classification head usually accomplishes the classification task through only one FC layer. This approach ignores the heterogeneity of features themselves. For example, shallow features may contain more detailed information, such as textures and edges, while deep features focus more on semantic information, such as the overall shape of objects. A single classification head finds it difficult to take into account the contributions of features at different levels simultaneously, which may lead to a bottleneck in classification performance.

The VGG and Swin Transformer Backbones selected for this study exhibit similar limitations. To address the aforementioned issues, this study proposes a Multi-Head Classifier architecture. Its core concept is to enhance the model's classification ability through hierarchical feature disentanglement and joint optimization. The theoretical basis of this method can be summarized in the following two points:

1. Divide-and-Conquer Strategy. The multi-head classifier adopts a divide-and-conquer approach by partitioning feature layers into multiple subsets (Heads) based on hierarchical levels or channels. Each head specializes in learning feature representations at specific scales, thereby reducing inter-feature interference. This design draws inspiration from divide-and-conquer algorithms, decomposing complex tasks into parallel subtasks. It enhances computational efficiency while reducing model complexity.

2. Feature Disentanglement and Complementarity. Different feature layers capture distinct levels of semantic information; shallow features contain local details while deep features encode global semantics. By assigning distinct feature subsets to separate classifi-

cation heads, the model avoids the limitations of a single-head representation and enhances category discrimination through fused multi-head outputs.

This work aims to propose a multi-head classifier for a multiple neural network. The improved architecture can enhance Top-k accuracy through hierarchical feature disentanglement and joint optimization, explore the contributions of different feature layers, and validate the improvements in model robustness and generalization capabilities achieved by the multi-head classifier.

5.2. Limitations Analysis

Firstly, our method exhibits a certain sensitivity to the backbone network architecture. As the experimental results in Section 4.3 show, while our approach achieves significant performance improvements on CNN-based backbones like ResNet and VGG, the gains are more limited on the Swin Transformer architecture. We hypothesize that this is because the self-attention mechanism in Transformer models is inherently a powerful, implicit tool for feature interaction and disentanglement. The explicit, strong disentanglement constraint we impose may, in some cases, conflict with or become redundant to the dynamic weight allocation of self-attention, thereby limiting further performance gains.

Secondly, the proposed multi-head structure entails a trade-off in computational efficiency. Compared to a traditional single-head classifier, our introduction of an independent classification head for each class inevitably increases the model's parameter count and computational load. This overhead becomes particularly pronounced in scenarios with an extremely large number of classes. Although this cost was acceptable for the datasets used in our experiments, it remains a practical factor to consider for real-world deployment.

Finally, enforcing strict class-level feature disentanglement may be suboptimal in certain contexts. Visual concepts in the real world are often not entirely independent but exhibit inherent co-occurrence or correlative relationships. As our method primarily focuses on learning discriminative features between classes, its current design may not explicitly model and leverage these beneficial inter-class associations.

6. Conclusions

The core innovation of this study lies in its priori interpretability, which serves as the fundamental design principle of the multi-head classifier. Starting from the feature layers of the backbone network, multiple structurally identical yet parameter-independent classification heads are precisely directed toward their respective class labels to collectively learn unified feature disentanglement capabilities—this constitutes a direct manifestation of priori interpretability. From model architecture and experimental design to the conceptual framework of our custom-built dataset, the concept of priori interpretability is consistently embedded throughout. Experimental results further demonstrate that the multi-head classifier, across diverse backbone networks and datasets, consistently exhibits the ability to learn feature disentanglement for multi-label image classification tasks. This validates the feasibility and scientific validity of priorily interpretable neural networks. Our research provides a novel perspective on neural network interpretability, further lifting the veil on the "black box".

Author Contributions: Conceptualization, Y.T. and Y.L.; Methodology, Z.L.; Software, Y.T.; Validation, Y.T.; Formal analysis, Z.L.; Investigation, Y.L.; Resources, Z.L. and Y.L.; Data curation, Y.D.; Writing—original draft preparation, Y.T.; Writing—review and editing, Y.L. and Z.L.; Visualization, Y.D.; Supervision, Y.L.; Project administration, Y.L.; Funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Project of Gansu (No. 24JRRA388).

Institutional Review Board Statement: Not applicable. This study does not involve human participants, animals, or biological samples.

Informed Consent Statement: Not applicable. This study does not involve human participants or patients.

Data Availability Statement: The data that support the findings of this study will be made openly available upon publication of this article.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

Grad-CAM	Gradient-weighted Class Activation Mapping
HL	Hamming Loss

References

1. Ayoobi, H.; Potyka, N.; Toni, F. ProtoArgNet: Interpretable image classification with super-prototypes and argumentation. In Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 1791–1799.
2. Chowdhary, C.L.; Acharjya, D.P. Segmentation and feature extraction in medical imaging: A systematic review. *Procedia Comput. Sci.* **2020**, *167*, 26–36. [CrossRef]
3. Lin, T.-Y.; RoyChowdhury, A.; Maji, S. Bilinear CNNs for fine-grained visual recognition. *arXiv* **2015**, arXiv:1504.07889.
4. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
5. Caruana, R. Multitask learning. *Mach. Learn.* **1997**, *28*, 41–75. [CrossRef]
6. Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2921–2929.
7. Lu, Z.; Lu, Y. Enhancing the reliability of image classification using the intrinsic features. *Knowl.-Based Syst.* **2023**, *263*, 110256. [CrossRef]
8. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European Conference on Computer Vision, Zürich, Switzerland, 6–12 September 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 818–833.
9. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems, Proceedings of the Conference and Workshop on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017*; NeurIPS Foundation: La Jolla, CA, USA, 2017; Volume 30.
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
12. Jain, S.; Wallace, B.C. Attention is not Explanation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1.
13. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You? Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
14. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *arXiv* **2017**, arXiv:1705.07874. [CrossRef]
15. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef] [PubMed]
16. Arafat, I.M.; Ferdous, S.M.S.; Quarles, J. Cybersickness-provoking virtual reality alters brain signals of persons with multiple sclerosis. In Proceedings of the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Tübingen/Reutlingen, Germany, 18–22 March 2018; pp. 1–120.
17. Dara, S.; Tumma, P. Feature extraction by using deep learning: A survey. In Proceedings of the 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 29–31 March 2018; IEEE: New York, NY, USA, 2018; pp. 1795–1801.

18. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems, Proceedings of the Conference and Workshop on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016*; NeurIPS Foundation: La Jolla, CA, USA, 2016; Volume 29.
19. Chen, X.; Kingma, D.P.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; Abbeel, P. Variational lossy autoencoder. *arXiv* **2016**, arXiv:1611.02731.
20. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
21. Tjoa, E.; Guan, C. New opportunities for the formal proof of computational real geometry? *arXiv* **2020**, arXiv:2004.04034. [CrossRef]
22. Arjovsky, M.; Bottou, L.; Gulrajani, I.; Lopez-Paz, D. Invariant risk minimization. *arXiv* **2019**, arXiv:1907.02893.
23. Yao, S.; Guan, R.; Huang, X.; Li, Z.; Sha, X.; Yue, Y.; Lim, E.G.; Seo, H.; Man, K.L.; Zhu, X.; et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *IEEE Trans. Intell. Veh.* **2023**, *9*, 2094–2128. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.