fire

Article

Advanced Multi-Label Fire Scene Image Classification via BiFormer, Domain-Adversarial Network and GCN

Yu Bai [1], Dan Wang [1,*], Qingliang Li [2], Taihui Liu [1] and Yuheng Ji [1]

[1] College of Computer Science and Technology, Bei Hua University, Jilin 132013, China
[2] College of Computer Science and Technology, Changchun Normal University, Changchun 130032, China
[*] Correspondence: 210280639zyhxq@gmail.com

Abstract: Detecting wildfires presents significant challenges due to the presence of various potential targets in fire imagery, such as smoke, vehicles, and people. To address these challenges, we propose a novel multi-label classification model based on BiFormer's feature extraction method, which constructs sparse region-indexing relations and performs feature extraction only in key regions, thereby facilitating more effective capture of flame characteristics. Additionally, we introduce a feature screening method based on a domain-adversarial neural network (DANN) to minimize misclassification by accurately determining feature domains. Furthermore, a feature discrimination method utilizing a Graph Convolutional Network (GCN) is proposed, enabling the model to capture label correlations more effectively and improve performance by constructing a label correlation matrix. This model enhances cross-domain generalization capability and improves recognition performance in fire scenarios. In the experimental phase, we developed a comprehensive dataset by integrating multiple fire-related public datasets, and conducted detailed comparison and ablation experiments. Results from the tenfold cross-validation demonstrate that the proposed model significantly improves recognition of multi-labeled images in fire scenarios. Compared with the baseline model, the mAP increased by 4.426%, CP by 4.14% and CF1 by 7.04%.

Keywords: wildfire detection; attention mechanisms; multi-label image recognition; transfer learning; data augmentation

check for updates

Citation: Bai, Y.; Wang, D.; Li, Q.; Liu, T.; Ji, Y. Advanced Multi-Label Fire Scene Image Classification via BiFormer, Domain-Adversarial Network and GCN. *Fire* **2024**, *7*, 322. https://doi.org/10.3390/fire7090322

Academic Editors: Darko Stipanicev and Giovanni Laneve

Received: 20 August 2024
Revised: 8 September 2024
Accepted: 11 September 2024
Published: 15 September 2024

1. Introduction

In recent years, wildfires have increased in both severity and frequency worldwide. These fires not only devastate infrastructure and result in casualties among both firefighters and civilians [1,2], but they also release large amounts of carbon dioxide, causing significant environmental harm [3,4]. To address these issues, computer vision methods have been applied to fire hazard identification and risk assessment, including applications such as fuel mapping [5], burned area detection [6], and wildfire detection [7]. However, the complexity of wildfire scenarios, which involve a wide range of fire-related factors, presents several challenges for computer vision-based wildfire detection. The main issues are outlined below.

First Challenge: A wildfire detection model based on transfer learning for ternary classification was proposed in [3], capable of detecting flames, smoke, and non-fire elements. However, this model only utilizes a CNN to process image features, which is insufficient for handling complex fire scenarios. To address this, we propose a feature extraction method based on BiFormer [8], which constructs sparse region-indexing relations and focuses on extracting features only in key regions. This allows the model to capture critical flame features for more effective subsequent processing.

Second Challenge: A multi-label prediction model based on transfer learning, as proposed in [4], identifies six labels. Despite some success in multi-label prediction [9], it does not account for essential objects like cars, people, and buildings—elements that may contribute to secondary fires in wildfire scenarios. Moreover, it does not specifically process these labels. To resolve this, we introduce a feature discrimination method based on GCN [10], which constructs a label association matrix to capture the correlation between labels, improving the model's ability to predict multiple labels in fire scenarios.

Third Challenge: Although elements such as clouds and the sun were included in the dataset in [4], no specific processing was performed to reduce the misclassification of these visually similar objects as fire [11]. To address this issue, we propose a feature screening method using a DANN [12,13], which divides the model's features into domains after extraction, effectively reducing the misclassification of similar elements.

Fourth Challenge: The limited availability of fire-related datasets hinders effective model training. While transfer learning alleviates some of the issues caused by insufficient data, it does not fully resolve the problem and may lead to overfitting, which can degrade model performance. Moreover, most fire datasets are restricted to just three labels, which do not reflect the complexity of real-world fire scenarios. To address this, we applied data augmentation and transfer learning techniques, while also integrating several publicly available datasets, including the Corsican Fire Database [14], VOC2012 [15], and the Korean Tourist Spot (KTS) dataset [16]. These datasets include essential fire-related elements such as cars, buildings, and people, as well as objects like clouds and the sun, which are prone to causing misclassification. The integrated dataset contains ten labels, including flames, wildfires, smoke, pedestrians, vehicle accidents, and more.

In this study, we propose a multi-label fire prediction model designed to better address complex fire scenarios. The effectiveness of the model is demonstrated through comparison and ablation experiments, with results visualized using class activation maps (CAM).

2. Related Work

Traditionally, wildfires have been detected primarily through human observation, using fire towers or detection cameras. However, these methods are prone to observer error and are limited by spatial and temporal constraints [17]. More recently, Maria Joao Sousa et al. proposed a transfer learning technique integrated with data augmentation, which was validated through tenfold cross-validation experiments [3]. Despite its improvements, this method performs only binary classification and fails to provide additional contextual information about the surrounding environment, highlighting the limitations of the dataset. Conversely, the authors in [4] introduced a similar model based on transfer learning that supports multi-label classification, offering predictions for multiple labels compared to its predecessor. While this approach achieves hexadecimal multi-label classification, it does not significantly improve the overall model performance.

During model validation, the color and texture of smoke often resemble other natural phenomena such as fog, clouds, and water vapor. Similarly, the color of flames can resemble the sun. As a result, algorithms based on smoke detection tend to generate a high rate of false positives, especially during nighttime detection [3,18]. To mitigate this issue, recent research has focused on incorporating these visually similar objects into the dataset. Consequently, it is essential to classify these confounding elements separately to improve detection accuracy. Traditional multi-label recognition presents several challenges. A common approach is to train independent binary classifiers for each label, but this method overlooks label relationships and can lead to exponential label growth as categories increase [19]. To address this issue, researchers have focused on capturing label dependencies. Wang et al. [20] employed recurrent neural networks (RNNs) to convert labels into embedded vectors, enabling the modeling of label correlations. Additionally, Wang et al. [21] incorporated a spatial transformer layer and Long Short-Term Memory (LSTM) units to capture label dependencies. Graph structures have proven more effective than the aforementioned learning methods for modeling label correlations. Li et al. [22] applied the Maximum Spanning Tree algorithm to create a tree-structured graph in label space, while Li et al. [23] generated image-dependent conditional label structures using the graphical Lasso framework. Lee et al. [24] utilized knowledge graphs to model multiple label relationships. These studies demonstrate the efficacy of graph-based methods, which is why this paper also employs Graph Convolutional Networks (GCNs) to capture and explore label interdependencies. GCNs propagate information across multiple labels, enabling the learning of interdependent classifiers for each image label. These classifiers leverage information from the label graph and apply it to the global image representation for final multi-label prediction. We also draw upon the work of Yaroslav Ganin et al. [12], who introduced a domain-adaptive representation learning method based on their DANN (domain-adversarial neural network) architecture. This architecture consists of three linear components: a label predictor, a domain classifier, and a feature extractor. It has demonstrated effectiveness in applying domain-adversarial learning. In this paper, we employ ResNet101 as the base network for feature extraction, which then feeds into a DANN's domain classifier and label predictor. Finally, the model integrates with a GCN to produce the final output. Experimental results confirm the efficiency of our proposed approach, which is fully compatible with end-to-end training.

3. Materials and Methods

In this section, we provide a detailed overview of the proposed model. First, we describe the reconstructed dataset, followed by an explanation of the model's key components. Lastly, we outline the architecture of the relevant network model used in this study.

3.1. Dataset

Current fire prediction datasets are not comprehensive and mostly consist of binary or ternary classified data, which are insufficient for training models capable of handling complex fire scenarios. To address this, we aimed to develop datasets containing ten categories, providing more detailed information about the surrounding scene to enable decision-makers to make more accurate judgments. In our study, we integrated datasets such as the Corsican Fire Database [14], Korean Tourist Spot (KTS) [16], VOC2012 [15], and various flame and smoke datasets downloaded from the web. The final dataset consists of 6213 images across ten categories: accidents, buildings, cars, clouds, flames, normal, people, smoke, sun, and wildfires.

The construction process was as follows: First, all data from the Corsican Fire Database were extracted, followed by the extraction of fire-related data, as well as data related to buildings, clouds, smoke, and the sun, from the Korean Tourist Spot dataset. From the VOC2012 dataset, images containing people and cars were primarily selected. After obtaining the required images, labels were manually annotated to classify each image into its respective category.

The Corsican Fire Database (CFDB) is an online image database that supports wildfire research by providing test datasets for the comparison of computer vision algorithms [10]. It is intended to be an evolving dataset and, as of October 2023, it consists of 1135 visible range images of wildfires. An example of this database is shown in Figure 1.

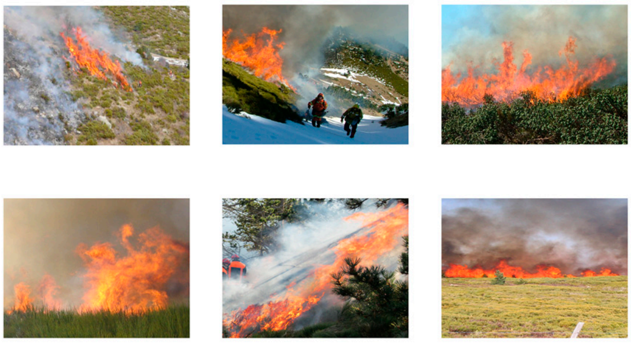

Figure 1. Rescaled samples of fire images from CFDB.

The Korean Tourist Spot (KTS) dataset contains numerous images of buildings constructed from wooden materials [16]. Additionally, images of clouds and the sun, which resemble early wildfire smoke and flames in color and shape, were included in the dataset due to their tendency to be incorrectly detected as wildfires. To reduce model misdetection, these images were incorporated into the dataset. A sample of the dataset is shown in Figure 2.

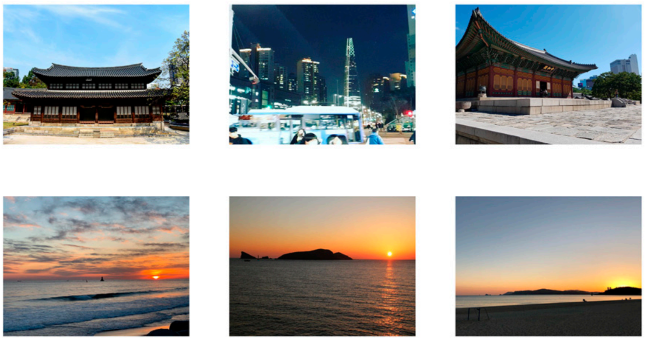

Figure 2. Rescaled samples of fire images from KT.

The VOC2012 dataset is a widely used standard for computer vision tasks, such as object detection, image segmentation, and recognition. It contains 20 categories of objects, including people, dogs, cats, cars, planes, bicycles, birds, chairs, bottles, and plants, with approximately one thousand images per category. For our study, we selected data from only two categories: people and cars. This selection was based on the frequent occurrence of these elements in fire scenes, where the presence of people is crucial for directing firefighting efforts, and cars may pose risks of secondary fires. Additionally, fire engines, which are also categorized as cars, play a vital role in fire suppression. A sample of the dataset is shown in Figure 3.

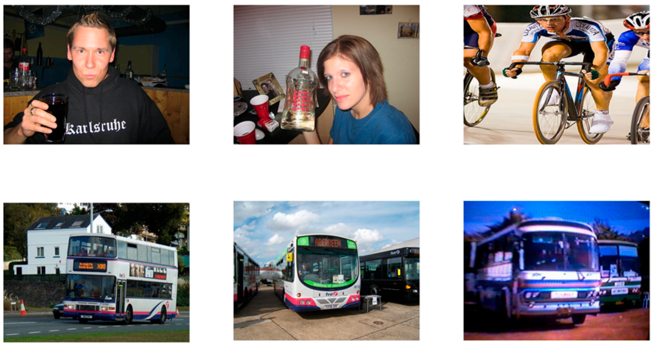

Figure 3. Rescaled samples of fire images from VOC2012.

The structural composition of our final integrated dataset is shown in Table 1. All instances were annotated according to the following classes: “Accident”, “Building”, “Car”, “Cloud”, “Flame”, “Non_Fire”, “Person”, “Smog”, “Sun”, and “Wildfire” (each class was abbreviated as “A”, “B”, “Ca”, “Cl”, “F”, “N”, “P”, “Sm”, “Su”, and “W”, respectively). We divided eighty per cent of the dataset into a training set and the remaining twenty per cent into a test set.

Table 1. Structure of the dataset.

	A	**B**	**Ca**	**Cl**	**F**	**N**	**P**	**Sm**	**Su**	**W**
Train	1178	870	1240	217	1200	2090	732	1675	154	400
Test	294	217	310	54	300	522	182	418	38	100

3.2. Basic Principle

The proposed model employs the ResNet101 network as the backbone for initial feature extraction. These extracted features are then passed to the BiFormer attention module, which enhances flame feature detection by constructing sparse region-indexing relations and focusing feature extraction on key regions. Subsequently, the processed features are input into the DANN for feature filtering, which performs domain segmentation to minimize misclassification. Finally, the GCN captures label correlations by constructing a label association matrix, enabling the final prediction. The network architecture is illustrated in Figure 4.

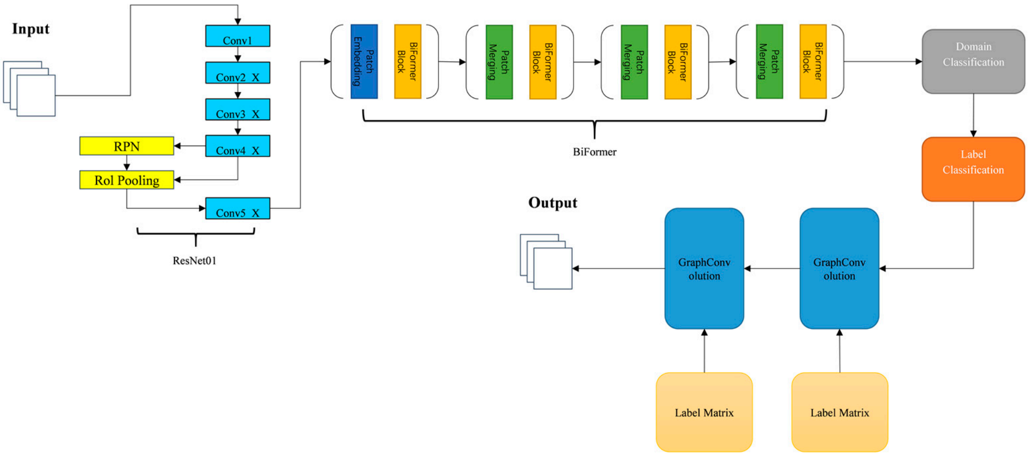

Figure 4. Model framework diagram.

3.3. Base Model

For this research, any CNN base model can be utilized to comprehend the characteristics of an image. ResNet-101 [25] was the preferred base model during our experiments, with DesNet121 [26] being employed in subsequent experiments for contrast. The image has a resolution of 448 × 448, and thus produces 2048 × 14 × 14 feature maps from the "conv 5x" layer. We subsequently performed a global pooling process on the output, having run it through a spatial and channel attention system. The product was then transmitted to the domain classification and label classification networks. Ultimately, the final outcomes are the consequence of the results received from the GCN being multiplied.

3.4. Graph Convolutional Networks

Graph Convolutional Networks (GCN) were first introduced in [25] for the task of semi-supervised classification.

Unlike standard convolution, which operates exclusively on local Euclidean structures in images, the goal of GCN is to acquire knowledge to learn a function f(·,·) on the graph G. This process requires input from the feature description $H^{l+1} \in R^{n\times d'}$ and the corresponding correlation matrix $A \in R^{n\times n}$ that signify n nodes and d dimensions of node features. After this input, the GCN will update the node features to $H^{l+1} \in R^{n\times d'}$. Each GCN layer can be represented as a nonlinear function.

$$H^{l+1} = f\left(H^{l}, A\right) \tag{1}$$

After applying the convolution operation described in reference [25], the function f(·,·) can be represented as the following:

$$H^{l+1} = h\left(\hat{A}H^{l}W^{l}\right) \tag{2}$$

where the transformation matrix to be learned is $W^l \in R^{d\times d'}$, the normalized correlation matrix $\hat{A} \in R^{\,n\times n}$, and the nonlinear operation $h(\cdot)$, which was performed by LeakyReLU in our experiments.

A GCN propagates information between nodes based on a correlation matrix, making the issue of how to build the matrix a crucial aspect of GCNs. This paper employs a data-driven approach to construct the correlation matrix A, defining the correlation between tags by mining co-occurrence patterns within the dataset [19].

We represent the relationship between label relevance using conditional probability, denoted by P(Lj | Li), which indicates the probability of label Lj appearing when label Li appears. As Figure 5 shows, P(Lj | Li) differs from P(Li | Lj), so the correlation matrix is asymmetric [19]. To create the correlation matrix, we count the number of instances of label pairs in the training set, resulting in a matrix M ∈ RC × C. C represents the total number of categories, and Mij represents the number of times Li and Lj occur together. Technical abbreviations are explained upon first usage. Using the label co-occurrence matrix, we can generate the conditional probability matrix according to Equation (3). In this equation, Ni represents the frequency of Li appearing in the training set and Pij = P(Lj | Li) indicates the likelihood of label Lj when label Li appears:

$$P_i = M_i / N_i \tag{3}$$

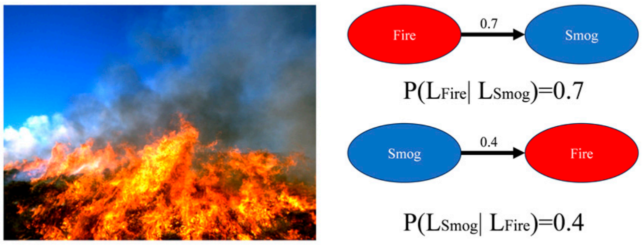

Figure 5. An example of the conditional probability relationship between two labels is provided. Typically, when the image contains "flame", there is a high likelihood that "smoke" is also present. However, if "smoke" is observed, "flame" may not necessarily be present.

However, there are two potential drawbacks with the aforementioned correlations. Firstly, the pattern of co-occurrences amongst labels and other labels may exhibit a long-tailed distribution, where some of the rarer co-occurrences could be considered extraneous. Secondly, the total amount of co-occurrences during the training and testing phases may not be equivalent, causing overfitting of the correlation matrix to the training set, thus leading to a loss of generalization [19]. Thus, we binarize correlation P by employing a threshold τ to remove noisy edges. This operation can be expressed as Equation (4).

$$A_{ij} = \begin{cases} 0, & P_{ij} < \tau \\ 1, & P_{ij} \geq \tau \end{cases} \tag{4}$$

Our primary aim in incorporating a GCN was to construct an efficient system for capturing the correlation between labels of multi-label images. Consequently, this can enhance the accuracy of multi-label predictions and reduce the potential misclassification due to similar labels.

3.5. Attention Mechanisms

It is widely recognized that attention plays a critical role in human perception. A key characteristic of the human visual system is its selective focus, where individuals process scenes through a sequence of fragmented glances, prioritizing salient features to efficiently comprehend visual information.

In this study, we incorporated the BiFormer Attention Block into the model to enhance its ability to capture accurate information while filtering out irrelevant data. The BiFormer architecture consists of four modules. Each module first downsamples the feature map using a convolutional layer, followed by feature extraction using the BiFormer Block. The structure of the BiFormer Block is illustrated in Figure 6. The BiFormer Block is composed of depthwise separable convolution, layer normalization, and bi-level routing attention, all connected by residuals. The bi-level routing attention operates in two stages: first, the feature map is projected through three fully connected layers to generate the query, key, and value matrices. The query and key matrices are then downsampled using average pooling at the regional level, and the correlation between them is calculated to filter out low-correlation key-value pairs. In the second stage, the undownsampled query matrix is used to compute attention only on the high-correlation key-value pairs derived from the first step. This dynamic sparse attention reduces computational complexity while effectively modeling long-range dependencies, thereby improving detection accuracy. The attention calculation is shown in Equation (5), where softmax is the activation function, C is the number of channels of the value matrix V, and T is the transpose operation. In the following experiments, we verify the effectiveness of incorporating the attention mechanism.

$$Attention(Q,K,V) = sotmax\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{5}$$

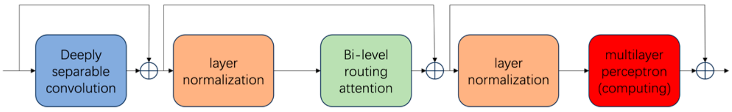
Deeply separable convolution
layer normalization
Bi-level routing attention
layer normalization
multilayer perceptron (computing)

Figure 6. BiFormer Block operational flow (⊕ Represents a residual connection).

3.6. Domain Classification and Label Classification Networks

In our experiments, we found that the results obtained from directly processing features extracted from the base model were unsatisfactory. The primary issue was that when making direct predictions, the model tended to misclassify visually similar labels. For example, the sun and flames, or smoke and clouds, share certain characteristics that can confuse the model, necessitating further feature refinement. In [12], the authors proposed a novel domain-adaptive method for feedforward neural networks, enabling large-scale training using annotated data in the source domain and unannotated data in the target domain. This approach is straightforward to implement in most deep learning frameworks by incorporating a simple gradient inversion layer. In our model, we added domain and label classification networks designed to filter features, extract the most relevant ones, and reduce the final error rate. Our experiments confirmed the effectiveness of this method. The structure of the domain and label classification networks is illustrated in Figure 7.

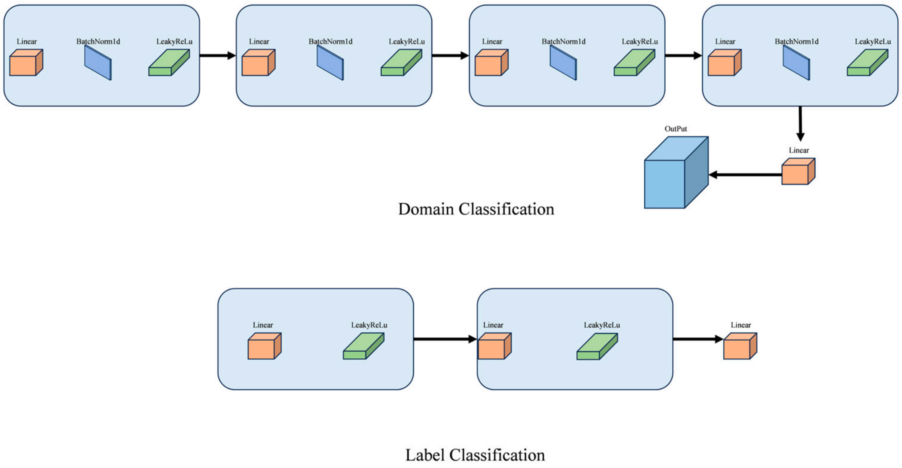

Figure 7. Domain classification and label classification network architecture.

3.7. Data Enhancement

Data augmentation is a widely used technique in Convolutional Neural Networks (CNNs) to increase the size of the training dataset and enhance model generalization. By applying various transformations and distortions to the original data, this technique generates new training samples, allowing the model to better adapt to data variations. Augmenting the dataset can significantly reduce error rates and improve model robustness. In this study, preprocessing operations, including MultiScale Crop and random horizontal flipping, were applied to ensure dataset diversity.

4. Results

In this section, we first describe the evaluation metrics and hyperparameter settings. The rest of this section presents the results of each model on the integrated dataset. This is followed by the corresponding visualization and finally a comprehensive analysis of the misclassifications and their possible causes.

4.1. Evaluation Metrics

In accordance with conventional settings [20], we present the mean per class precision (CP), recall (CR), F1 (CF1), mean overall precision (OP), recall (OR), and F1 (OF1) for performance evaluation. Additionally, we calculate and report the mean accuracy (mAP).

4.2. Hyperparameter Setting

This study compares various models through tenfold cross-validation and ablation experiments, evaluating their performance on reconstructed datasets. The hyperparameters for each model are detailed in Table 2. All models were trained for 100 epochs.

Table 2. Parameters for each model.

	DANN-GCN	ResNet101-GCN	ResNet101-DA-GCN	DesNet121-GCN
τ	0.41	0.41	0.41	0.41
Batch size	32	32	32	32
Learning rate	0.1	0.1	0.1	0.1
Optimizer	SGD	SGD	SGD	SGD

4.3. Experimental Results

4.3.1. Comparison with State-of-the-Art Technology

The integrated dataset consists of 10 categories and 6213 images. We allocated 80% of the images to the training set and the remaining 20% to the validation set. To augment the data, input images were randomly cropped and resized to 448 × 448 during training, and randomly flipped horizontally. To validate the proposed method, we used DenseNet121 as the backbone network for feature extraction. The attention mechanism, along with domain and label classification networks, were incorporated, and the model was further integrated with the GCN. The final experimental results, averaged over five trials, are shown in Table 3. The proposed model ranked first in five of the seven metrics and second in the remaining two. However, the model's performance slightly decreased after adding the GCN, likely due to insufficient feature extraction, which hindered the GCN from accurately constructing the label matrix. After incorporating the domain-adversarial network, both ResNet101 and DenseNet121 exhibited significant performance improvements, surpassing their respective base models.

Table 3. Compare the performance scores of each model.

	mAP	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
DANN-GCN	70.0914	0.5938	0.5318	0.5546	0.8500	0.6628	0.7348
ResNet101	93.606	0.9064	0.8348	0.8634	0.9122	0.8344	0.866
ResNet101-GCN	89.4530	0.8696	0.7818	0.8106	0.8760	0.7910	0.8174
ResNet101-DA-GCN	**93.879**	**0.911**	**0.868**	**0.881**	**0.926**	**0.891**	**0.899**
DesNet121	90.382	0.882	0.863	0.864	0.903	0.88	0.879
DesNet121-GCN	91.6434	0.8872	0.8212	0.8428	0.8852	0.819	0.841
DesNet121-DA-GCN	92.7438	0.8960	0.8216	0.8514	0.9044	0.8292	0.8614

4.3.2. Tenfold Cross-Validation

To enhance generalization and mitigate the risk of overfitting, this study employs tenfold cross-validation rather than a single train-validation split. The results, presented in Table 4, demonstrate that the model performs optimally in tenfold cross-validation, with most performance metrics ranking in the top three. The integration of the GCN with the ResNet101 model did not lead to as significant an improvement as observed with DenseNet121. This discrepancy may be attributed to differences in the underlying model architectures. Nonetheless, the overall performance of the model improved with the incorporation of the domain-adversarial network.

Table 4. Tenfold cross-validation results.

	mAP	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
DANN-GCN	65.295	0.806	0.527	0.536	0.806	0.746	0.74
ResNet101	87.379	0.859	0.744	0.759	**0.902**	0.863	**0.874**
ResNet101-GCN	79.569	0.784	0.772	0.738	0.834	0.866	0.787
ResNet101-DA-GCN	**88.987**	**0.877**	0.838	**0.843**	0.889	0.865	0.868
DesNet121	73.549	0.471	0.737	0.555	0.42	0.722	0.514
DesNet121-GCN	86.346	**0.865**	**0.843**	0.828	0.892	**0.871**	0.859
DesNet121-DA-GCN	88.144	0.848	0.758	0.756	0.892	0.869	0.871

4.3.3. Visualization of Results

To evaluate the effectiveness of the added attention mechanism, we visualized the results using the GAP-CAM method [27]. We compared the visualizations before and after applying the attention mechanism. The results are presented in Figure 8.

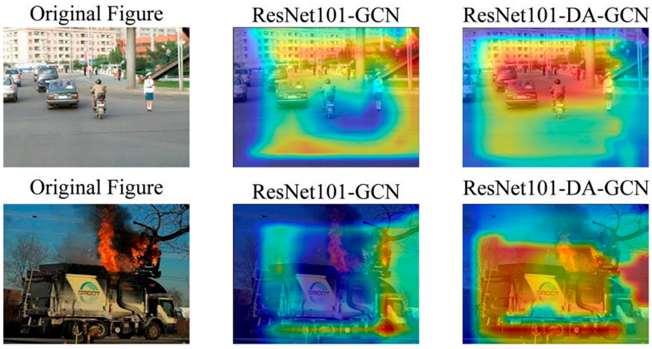

Figure 8. Visualization of results (where red represents a higher level of concern).

The figure above illustrates that the baseline model fails to focus on several important features and often neglects others. In contrast, our proposed model effectively emphasizes most of the critical features. To evaluate the effectiveness of our model in learning improved image representations, we conducted an image classification experiment. A Multi-Label Graph Convolutional Network (MLGCN) [19] was used as the baseline for comparison.

We presented three images for classification, and the results are depicted in Figure 9. Our model's performance significantly surpasses that of the baseline; as shown in Figure 9, the labels predicted by our method closely match the true labels of the query images. These results indicate that our model not only effectively captures label dependencies to enhance classifier performance but also benefits from advanced image representation learning in multi-label recognition.

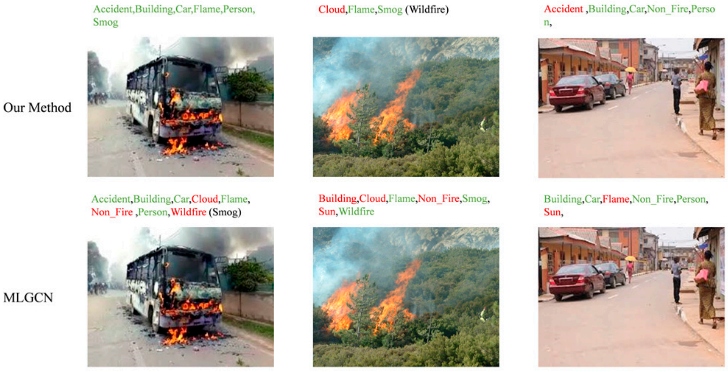
Accident,Building,Car,Flame,Person,
Smog
Cloud,Flame,Smog (Wildfire)
Accident ,Building,Car,Non_Fire,Perso
n,
Our Method
Accident,Building,Car,Cloud,Flame,
Non_Fire ,Person,Wildfire (Smog)
Building,Cloud,Flame,Non_Fire,Smog,
Sun,Wildfire
Building,Car,Flame,Non_Fire,Person,
Sun,
MLGCN

Figure 9. Visualization of multi-label classification results (where green means the prediction is correct and red means the prediction is incorrect).

4.4. Ablation Studies

In this section, we investigate ablation from two different aspects, including the effectiveness of our proposed plug-in and the effect of τ in the binarization of the correlation matrix.

4.4.1. The Effect of Different Thresholds τ

We change the value of the threshold τ in Expression 4, as shown in Figure 10. Note that we only considered the range of τ between 0.35 and 0.45, because we found that the results were less satisfactory for values in other ranges, so we did not include them in the results again. As shown in the figure, when edges with low probability (i.e., noisy edges) are filtered out, the accuracy of multi-label recognition improves. However, if too many edges are filtered out, the accuracy decreases because the relevant neighbors are also ignored. According to our experiments, the optimal values of τ are all 0.41.

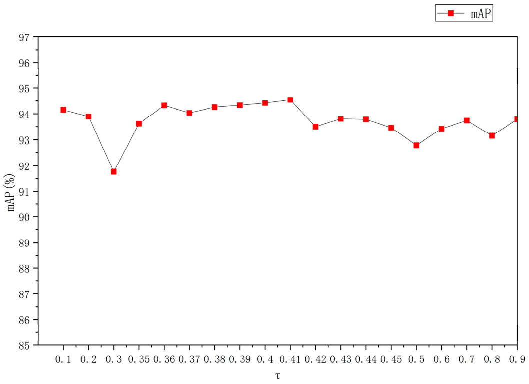

Figure 10. Accuracy comparisons with different values of τ.

4.4.2. Impact of Plug-Ins

We experimented with the domain and label classification networks and the attention mechanism on ResNet101 [25] and DesNet121 [26], respectively. The final results are shown in Table 5, and all show some improvement after the addition of our plug-ins.

Table 5. Comparison results of model performance after adding the plug-in.

	mAP	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
ResNet101	88.2316	0.8765	0.7923	0.8217	0.8973	0.8012	0.8191
ResNet101-DA-GCN	**93.879**	**0.911**	**0.868**	**0.881**	**0.926**	**0.891**	**0.899**
DesNet121	90.382	0.8673	**0.8346**	0.8287	0.8915	0.8101	0.8327
DesNet121-DA-GCN	**92.7438**	**0.8960**	0.8216	**0.8514**	**0.9044**	**0.8292**	**0.8614**

We initially validated the BiFormer attention module independently, which enhances flame feature detection by constructing sparse region-indexing relations and focusing feature extraction on critical areas. The BiFormer module was integrated into both ResNet101 [25] and DenseNet121 [26] for experimentation. As shown in Table 6, the results demonstrate improved model performance after incorporating the attention module.

Table 6. Comparison results of model performance after adding the plug-in.

	mAP	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
ResNet101	88.2316	0.8765	0.7923	0.8217	0.8973	0.8012	0.8191
ResNet101-BiFormer	**90.1756**	**0.8861**	**0.8012**	**0.8431**	**0.9013**	**0.8762**	**0.8873**
DesNet121	90.382	**0.8673**	**0.8346**	0.8287	0.8915	0.8101	0.8327
DesNet121-BiFormer	**91.1145**	0.8456	0.8091	**0.8378**	**0.8987**	**0.8193**	**0.8572**

Next, we independently validated the GCN, which captures label correlations by constructing a label association matrix, enabling more accurate multi-label prediction. The GCN module was integrated into both ResNet101 [25] and DenseNet121 [26] for experimentation. As shown in Table 7, the results demonstrate improved model performance after incorporating the GCN.

Table 7. Comparison results of model performance after adding the plug-in.

	mAP	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
ResNet101	88.2316	0.8765	0.7923	0.8217	0.8973	0.8012	0.8191
ResNet101-GCN	**91.1373**	**0.9014**	**0.8257**	**0.8312**	**0.9139**	**0.8619**	**0.8534**
DesNet121	90.382	0.8673	**0.8346**	0.8287	**0.8915**	0.8101	0.8327
DesNet121-GCN	**91.6487**	**0.8712**	0.8127	**0.8349**	0.8812	**0.8137**	**0.8479**

Finally, we independently validated the DANN, which filters features by dividing them into domains to reduce model misclassification. The DANN module was integrated into both ResNet101 [25] and DenseNet121 [26] for experimentation. As shown in Table 8, the results demonstrate improved model performance after incorporating the DANN.

Table 8. Comparison results of model performance after adding the plug-in.

	mAP	**CP**	**CR**	**CF1**	**OP**	**OR**	**OF1**
ResNet101	88.2316	0.8765	0.7923	0.8217	0.8973	0.8012	0.8191
ResNet101-DA	**90.5731**	**0.8801**	**0.8428**	**0.8356**	**0.9018**	**0.8427**	**0.8361**
DesNet121	90.382	**0.8673**	**0.8346**	0.8287	0.8915	0.8101	0.8327
DesNet121-DA	**90.9147**	0.8517	0.8216	**0.8417**	**0.8992**	**0.8184**	**0.8579**

The experiments indicate that while each individual module enhances model performance, the combined use of all modules yields a more significant improvement. Therefore, we recommend using the modules in combination.

4.4.3. False Positives

The reporting rate significantly affects the reliability of fire detection classifiers and their applicability to real-world scenarios [3]. This paper specifically addresses the challenge of distinguishing between sun and clouds. The color and shape of the sun closely resemble those of flames, while clouds and smoke are often comparable in appearance. Consequently, these similarities frequently lead to inaccurate detections. To address this issue, this paper incorporates images of the sun and clouds into the reconstructed dataset. This approach aims to mitigate potential errors during the validation stage and to develop more robust models for detecting wildfire-like images. The fire prediction scores for all images shown in Figure 11 should be below 0.5. While our models accurately detect the presence of fire, the MLGCN model [19] occasionally misclassifies the sun as a flame. This tendency is evident from the results, where the MLGCN often mistakes both flames and the sun for the sun. Figure 12 demonstrates that all smoke prediction scores are below 0.5, and our proposed model exhibits significantly fewer misclassifications compared to the MLGCN.

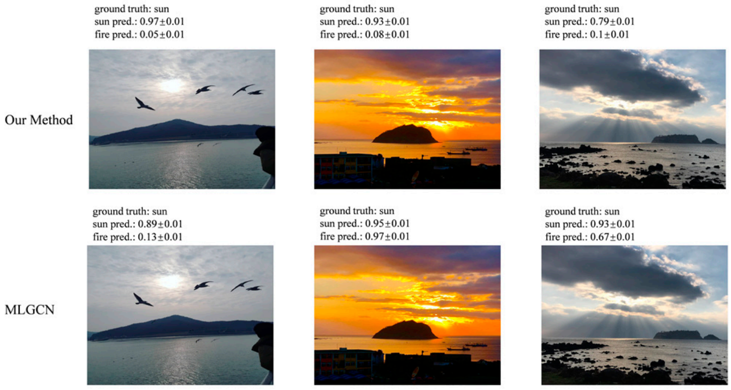

Figure 11. Example of predicting sun.

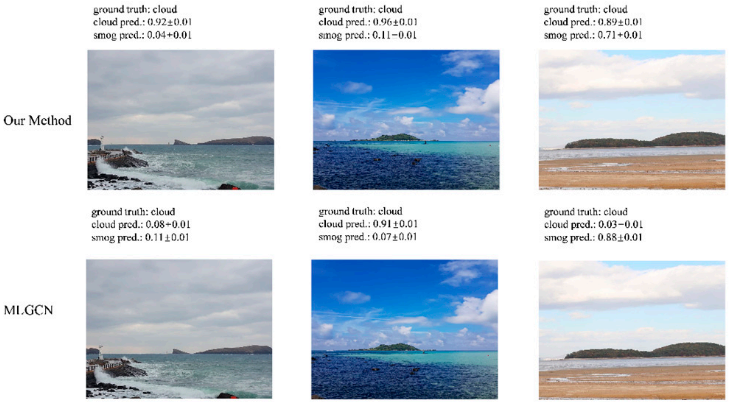

Figure 12. Example of predicting clouds.

To ensure the validity of the experiment, we performed a reverse comparison by presenting the model with images of flames and smoke to assess its ability to mistakenly predict the sun or clouds. The results are illustrated in Figure 13. Our recent findings indicate that while both models exhibit misclassifications, the proposed model significantly outperforms the MLGCN. This result underscores the effectiveness of incorporating both the domain classification network and the label classification network.

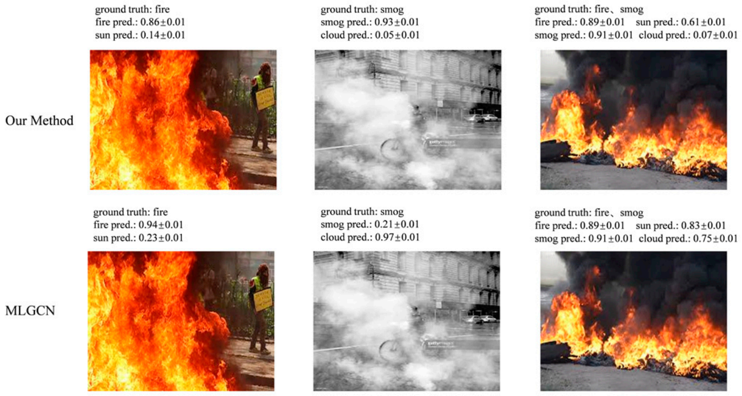

Figure 13. Example of predicting fire and smog.

5. Discussion

In this study, we propose a model that combines a GCN, a DANN, and the BiFormer attention mechanism. The experimental results demonstrate the effectiveness of this multi-label image recognition model, showing significant improvements in key performance metrics such as mAP, CP, CR, and F1 scores compared to the baseline. These results indicate that integrating a GCN for learning label correlations and applying domain-adversarial training to mitigate domain-specific biases greatly enhances the model's ability to accurately classify multiple labels in complex fire scenarios. Our model surpasses several state-of-the-art methods, particularly in scenarios involving multiple objects (e.g., smoke, vehicles, and people). The BiFormer attention mechanism, designed to capture long-range contextual dependencies, effectively improves the model's focus on the most relevant features of fire scenes. This is crucial for reducing false alarms, which are common in fire detection tasks due to visual similarities between fire-related objects and other environmental factors, such as clouds or sunlight.

A major challenge in this research is the limited availability of labeled fire-related datasets. Although we addressed this by integrating multiple public datasets, the combined datasets may still not fully represent the diversity of real fire scenarios, particularly regarding variations in lighting, weather conditions, and geography. Another challenge is the occasional misclassification of visually similar objects (e.g., the sun and flames, or clouds and smoke). While our model reduces such errors compared to the baseline, there is still room for improvement in distinguishing these subtle differences, especially when fire-related objects are occluded or only partially visible.

Despite these improvements, performance gains were not consistent across all metrics and models. For instance, adding the GCN to the ResNet101 model resulted in less improvement compared to the DenseNet121 model, indicating that the underlying architecture significantly influences how well the GCN leverages label correlation. This highlights the need for further investigation into the compatibility of GCNs with various backbone architectures.

Future work could explore other graph-based models or hybrid architectures that combine GCNs with advanced techniques such as transformer, potentially enhancing the model's ability to capture complex label dependencies. Additionally, implementing real-time testing in dynamic, real-world fire detection systems will be crucial for evaluating the model's robustness and practical applicability.

6. Conclusions

Previous computer vision-based fire management frameworks were limited to binary or ternary classification. However, in disaster response scenarios, decision-makers must consider multiple factors such as surrounding buildings, vehicles, and people to prioritize firefighting actions. Current fire datasets, however, do not encompass a sufficiently wide range of information, potentially leading to models that yield suboptimal results.

Additionally, the performance of existing fire prediction models in handling multi-label classification is inadequate for providing accurate information to decision-makers. To address these challenges, we integrated several fire-related datasets and constructed a new dataset containing ten labels. We also propose a novel model based on the MLGCN framework, which enhances the model's ability to capture label correlations by incorporating a GCN and the BiFormer attention mechanism. The model employs both a domain classification network and a label classification network for feature filtering. The proposed model was validated using the integrated dataset, and the results were visualized using Grad-CAM. Final metrics such as the mAP, CP, and CR surpassed the baseline model by over four percent, confirming the advantages of our approach. Currently, transformer models not only demonstrate exceptional performance in natural language processing but also exhibit unique capabilities in image processing. In the future, we will explore embedding a transformer model to further improve contextual relevance in image analysis.

Author Contributions: Conceptualization, Y.B.; writing—original draft preparation, Y.B.; visualization, Y.J.; supervision, D.W., Q.L.; funding acquisition, D.W., Q.L. and T.L. All authors have read and agreed to the published version of the manuscript.

Funding: The study was funded by the Natural Science Foundation of China, grant number 42275155; 42105144; 41975122; U1811464, in part by the Jilin Province Science and Technology Development Program, grant number 20230101370JC, the Jilin Provincial Department of Science and Technology's Natural Science Foundation project (YDZJ202201ZYTS558), Jilin Provincial Education Science Planning General Project (GH20276), Jilin Provincial Key Research and Development project (20200404223YY), Jilin Provincial Science and Technology Department project (YDZJ202201ZYTS605), and the General Project of Graduate Innovation Program at Beihua University ([2023]054).

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author, Dr. Dan Wang, upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Talaat, F.M.; ZainEldin, H. An improved fire detection approach based on YOLO-v8 for smart cities. *Neural Comput. Appl.* **2023**, *35*, 20939–20954. [CrossRef]
2. Zhao, L.; Zhi, L.; Zhao, C.; Zheng, W. Fire-YOLO: A small target object detection method for fire inspection. *Sustainability* **2022**, *14*, 4930. [CrossRef]
3. Sousa, M.J.; Moutinho, A.; Almeida, M. Wildfire detection using transfer learning on augmented datasets. *Expert Syst. Appl.* **2020**, *142*, 112975. [CrossRef]
4. Park, M.; Tran, D.Q.; Lee, S.; Park, S. Multilabel Image Classification with Deep Transfer Learning for Decision Support on Wildfire Response. *Remote Sens.* **2021**, *13*, 3985. [CrossRef]
5. Boschetti, L.; Roy, D.P.; Justice, C.O.; Humber, M.L. MODIS–Landsat fusion for large area 30 m burned area mapping. *Remote Sens. Environ.* **2015**, *161*, 27–42. [CrossRef]
6. Alonso-Benito, A.; Hernández-Leal, P.A.; Arbeló, M.; González-Calvo, A.; Moreno-Ruiz, J.A.; García-Lázaro, J.R. Satellite image based methods for fuels maps updating. In Proceedings of the SPIE, Edinburgh, UK, 25 October 2016. [CrossRef]
7. Chen, Y.; Liu, Z. Aerial Images-Based forest fire detection for firefighting using optical remote sensing techniques and unmanned aerial vehicles. *J. Intell. Robot. Syst.* **2017**, *88*, 635–654. [CrossRef]
8. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. BiFormer: Vision Transformer with Bi-Level Routing Attention. *arXiv* **2023**. [CrossRef]
9. Zhu, X.; Liu, J.; Liu, W.; Ge, J.; Liu, B.; Cao, J. Scene-Aware Label Graph Learning for Multi-Label Image Classification. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 1473–1482. [CrossRef]
10. Kipf, T.N.; Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv* **2016**. [CrossRef]
11. Wu, S.; Zhang, L. Using popular object detection methods for real-time forest fire detection. In Proceedings of the 2018 11th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 8–9 December 2018; IEEE: Piscataway, NJ, USA, 2018; Volume 1, pp. 280–284.
12. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial training of neural networks. In *Advances in Computer Vision and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 189–209. [CrossRef]

13. Zheng, J.; Wen, Y.; Chen, M.; Yuan, S.; Li, W.; Zhao, Y.; Wu, W.; Zhang, L.; Dong, R.; Fu, H. Open-set domain adaptation for scene classification using multi-adversarial learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *208*, 245–260. [CrossRef]
14. Toulouse, T.; Rossi, L.; Campana, A.; Çelik, T.; Akhloufi, M.A. Computer vision for wildfire research: An evolving image dataset for processing and analysis. *Fire Saf. J.* **2017**, *92*, 188–194. [CrossRef]
15. Everingham, M.; Eslami, S.M.A.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes Challenge: A Retrospective. *Int. J. Comput. Vis.* **2014**, *111*, 98–136. [CrossRef]
16. Jeong, C.; Jang, S.; Na, S.; Kim, J.T. Korean tourist spot Multi-Modal dataset for deep learning applications. *Data* **2019**, *4*, 139. [CrossRef]
17. Jain, P.; Coogan, S.C.P.; Subramanian, S.G.; Crowley, M.; Taylor, S.; Flannigan, M.D. A review of machine learning applications in wildfire science and management. *Environ. Rev.* **2020**, *28*, 478–505. [CrossRef]
18. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S. Convolutional neural networks based fire detection in surveillance videos. *IEEE Access* **2018**, *6*, 18174–18183. [CrossRef]
19. Chen, Z. Multi-Label Image Recognition with Graph Convolutional Networks. *arXiv* **2019**. [CrossRef]
20. Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; Xu, W. CNN-RNN: A unified framework for multi-label image classification. *arXiv* **2016**. [CrossRef]
21. Wang, Z.; Chen, T.; Li, G.; Xu, R.; Li, L. Multi-label image recognition by recurrently discovering attentional regions. *arXiv* **2017**. [CrossRef]
22. Li, X.; Zhao, F.; Guo, Y. Multi-Label Image Classification with a Probabilistic Label Enhancement Model. ResearchGate. 2014. Available online: https://www.researchgate.net/publication/287511295_Multi-label_image_classification_with_a_probabilistic_label_enhancement_model (accessed on 23 July 2014).
23. Li, Q. Conditional Graphical Lasso for Multi-Label Image Classification. 2016. Available online: https://www.semanticscholar.org/paper/Conditional-Graphical-Lasso-for-Multi-label-Image-Li-Qiao/b1d100cb3f1b39e2d9a28a941aa0d2999fa51fd2?utm_source=direct_link (accessed on 12 December 2016).
24. Lee, C.; Fang, W.; Yeh, C.; Wang, Y.F. Multi-Label Zero-Shot Learning with Structured Knowledge Graphs. *arXiv* **2017**. [CrossRef]
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *arXiv* **2015**. [CrossRef]
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. *arXiv* **2016**. [CrossRef]
27. Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. *arXiv* **2015**. [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.