Objection Detection Domain Generalization Research Based On Faster-Rcnn

Luyao Tang(23320221154300) [1] ,Mingbo Li(23020221154094) [1] ,Lidong Cheng(36920221153074) [1] ,Chenyu Ma(23320221154297) [2] ,Zhouxiang Xia(36920221153125) [1]

[1] Deep Learning Course Institute of Artificial Intelligence Class
[2] Deep Learning Course School of Informatics Class
lytang@stu.xmu.edu.cn, limingbo@stu.xmu.edu.cn, orange253@163.com, 23320221154297@stu.xmu.edu.cn,
zhouxiangxia@stu.xmu.edu.cn

Abstract

The domain generalization scheme aims to learn a model with strong generalization ability for multiple fields with different data distribution, so as to obtain better results in the unknown test dataset. Among them, most of the existing work of domain generalization focuses on the evaluation on the image classification datasets. For the more challenging object recognition task, the current research progress is very little.Aiming at the problems that the data collected from the actual application scene in the target detection task and the data used in the model training do not meet the independent and identically distributed criterion, which leads to the decline of the classification accuracy of the object detection model, the increase of the missed detection rate of the detection frame, the poor generalization ability and so on, this dissertation carries out the research on the domain generalization algorithm of the object detection based on machine learning, It is proposed to fully extract the cross domain information through the domain label and build a three-level graph convolution network at the pixel level, instance level and domain level. The loss function of the graph convolution network guides the model to build a cross domain generalized centroid, so that the model focuses on the domain invariant information in the process of object detection, strengthens the connection of different features of similar objects in the graph relationship network, and shortens the distance between similar objects in each domain in the feature space, To improve the accuracy of model detection. The project is verified on the dataset for the domain generalization task of object detection, which is collected and labeled by ourselves, and has four different data distributions. The result is better than the existing hybrid training algorithm, and has better generalization effect on the more abstract target domain, which greatly improves the ability of the model to be extended to the unknown domain.

lntroduction

Research Background

Target detection is one of the most widely concerned tasks in the field of machine vision. Currently, deep learning(Krizhevsky, Sutskever, and Hinton 2012) has become the mainstream solution in detection tasks. However, most current deep learning work still requires a large number of labeled samples as training data, and it is assumed that the training samples and test samples follow the requirements of Independent and Identical Distribution (IID)(Liu 2017). In practical applications, this requirement is difficult to meet. For example, in the target detection task, due to the influence of image acquisition equipment, illumination, weather, etc., it is difficult for data from different sources to satisfy the assumption that training data and test data are independent and identically distributed, and domain shift phenomenon(Ganin et al. 2016) appears. In turn, the target detection model that performed well on the training data has problems such as decreased target classification accuracy and inaccurate detection frame positions. In addition, due to the complexity and large amount of parameters of the deep learning model, it is easy to have overfitting problem during the training process, so it is more sensitive to domain shift. How to effectively use labeled data and reduce the waste of data resources has become one of the key issues in computer vision research. At present, the transfer learning theory represented by domain generalization and domain adaptation has become one of the mainstream methods to solve non-IID problems. At present, the domain adaptation theory has been used in target detection(Chen et al. 2020, 2021), but the domain adaptation algorithm needs to use unlabeled test data in the training process. Such target domain unlabeled data is not easy to obtain in practical application scenarios. Therefore, compared with domain adaptation, domain generalization is more suitable for practical applications because it does not need to use target domain data in the training process, and only uses labeled multi-source domain data to train a model with good generalization performance.

Based on the above background, this paper proposes a novel domain generalization target recognition model based on graph convolutional network, and collects and labels artificially simulated domain differences, and has four different data distributions for target detection domain generalization tasks. Verification is carried out on the data set to solve the problem that the training data in the target detection task and the data collected in the actual application scene do not meet the independent and identical distribution criterion, which leads to the decline of the model generalization ability, and improve the ability of the model to generalize to unknown domains.

Research Progress

Domain Adaptation With Object Detection The premise of the domain adaptation algorithm is that when the task conditions are the same, the data distribution of the source domain sample data set of the training set and the target domain sample data set of the test set do not satisfy the independent and identical distribution setting. It is committed to improving the generalization performance of the prediction model trained in the source domain to the target domain, mainly by reducing the domain-related features extracted by the model in the source domain data set to solve the domain shift problem.

The mainstream domain-adaptive target detection method, there are three main methods to solve the domain shift problem: fine-tuning based domain-adaptive target detection, source domain data to generate pseudo-labels for fine-tuning (Khodabandeh et al. 2019) or migration to real scenes via synthetic datasets(Cai et al. 2019). Domain-adapted target detection based on semantic alignment, by comparing the feature differences at different levels between different domains, guides the classifier to perform semantic alignment (Chen et al. 2018), and learns domain-invariant features (Zhu et al. 2019). Based on reconstruction domain adaptive target detection, the source domain data or target domain data are reconstructed to improve the model performance within a specific feature distribution(Arruda et al. 2019; Lin 2019). At present, better performance can be achieved by mixing multiple domain adaptive target detection methods, which mainly enhance local discriminability (Chen et al. 2020) by adversarial training and capturing potential complementary effects between global context information, or for samples with scarce Classes and variable samples are assigned greater weights, and features are forced to be aligned in cross-domain samples(Chen et al. 2021), and instance-invariant features are implicitly learned by exploring the natural features of unlabeled target domain data and training data, so that the model is in contact with After the target domain data can quickly match the features of the labeled training data, and then achieve the purpose of domain adaptation.

Domain Generalization With Object Detection Domain generalization cannot acquire any information of target domain data during training, so it is more challenging for domain generalization methods to extract generalized, transferable features from source domain samples. Domain generalization aims to learn a system that can maintain uniform and good performance across multiple different data distributions. Existing domain generalization methods can be divided into three categories (Wang et al. 2021): one is data manipulation, which increases the diversity of data through the enhancement and generation of training data. The second is representation learning, that is, domain-invariant representation learning, which is similar to domain adaptation. The purpose is to make the model adapt well to different fields. The learning of domain-invariant features mainly includes: Kernel methods, explicit feature alignment, domain adversarial training, and invariant risk minimization. The third is the learning strategy, which introduces mature learning methods in machine learning into cross-domain training, mainly based on ensemble learning and meta-learning methods to make the model more generalizable. In addition, self-supervised training can also be used in domain generalization.

At present, domain generalization has not been introduced into the target detection task, because without contact with the target domain data, the model has a strict limit on the distribution range of the feature space where the object can be detected. In the area of the space, the performance of the model is degraded, and it is very prone to false detection and missed detection. In the target detection task, for the situation where the category domain is invisible, that is, the zero-sample problem, a feature synthesizer can be constructed to generate domain features and corresponding domain labels (Huang et al. 2022) for unseen classes, but it can only be used in a certain To a certain extent, it solves the problem of false detection, but there is still a distance from practical application. In view of the fact that the distribution of the target domain closer to real life is invisible, the current study is how to generalize the model to a wider feature space on the basis of only contacting limited source domain data, so that the target detection model is applicable to a variety of Complex scenes are the most difficult point in the generalization task of target detection domain.

Main Content and Contributions

Aiming at the problems that traditional target recognition algorithms have limited training tasks, insufficient generalization experience, and insufficient correlation mining within data sets, this paper explores the domain generalization target recognition task. First of all, this model adopts a cross-domain training mode, which enriches the type and quantity of tasks by using the labels of each domain. The distribution of target domains in tasks is no longer limited to a single domain distribution, which is more in line with the mixed data sources and changeable data sources in actual application scenarios. imaging conditions. Secondly, this model introduces a learnable graph convolutional neural network to construct topological relationships at three levels of pixels, instances, and domains, and introduces cosine similarity into the construction of an adjacency matrix. Thirdly, this model adopts a learning and training method that is less related to the feature extractor for parameter updating, giving full play to the guiding role of the learnable cross-domain generalization centroid in the parameter updating process. The method proposed in this paper exposes the model to a wider range of domain shift scenarios by modifying the settings of the domain generalization task, enriches the generalization experience, optimizes the feature space, and thus improves the generalization ability of the model. The experimental results show that the classification accuracy and detection accuracy of the algorithm proposed in this paper exceed the existing hybrid training model in the target detection task, and it has a better generalization effect on the data set with a larger domain shift, which greatly improves It improves the ability of the model to generalize to unknown domains.

Basic Theory

In recent years, deep learning has performed well in the fields of computer vision, medical image processing, and natural language processing. Among them, Convolutional Neural Network (CNN), with its powerful feature extraction capabilities in the field of digital images, has produced far-reaching academic influence and huge commercial value in academic research and industry. However, the traditional convolutional neural network can only deal with Euclidean space data such as digital images and speech, and the data in these fields has translation invariance. In order to speed up the operation and reduce the amount of parameters, we can use this feature to define a globally shared convolution kernel in the input data space, thereby defining a convolutional neural network. Since the past five years, graph data has been introduced into the deep learning model (Kipf and Welling 2016). Data structures such as graphs can express real data in real life more naturally and intuitively, such as traffic routes and social network relationships. Wait. Different from digital image or speech data, the local structure of each node in graph data is different, which makes the translation invariance no longer satisfy (Shuman et al. 2013), and the learning method of graph convolution can be used to purposefully use the graph structure to describe Adjacent nodes, and information aggregation, for tasks with obvious differences in data distribution, the information exchange between nodes in the topology can achieve the purpose of improving the generalization ability of the model. This chapter will introduce the theoretical knowledge of target recognition network and graph convolutional network.

Theory of Object Detection

Overview of Object Detection Neural Network The three major tasks in the field of computer vision are classification, detection, and segmentation. Image classification models are used to match image content to a single category, using annotation information to artificially specify categories of interest. Most of the pictures in the real world are not close-ups of a single object. It is not accurate and practical to assign a single label to the image. For the situation where there are many types of objects in the picture and all of them need to be assigned labels and determine their positions, a target detection model is required. The target detection model can identify multiple objects in a frame captured by a picture or a video, and can locate different objects, that is, give a bounding box.

Target detection is essentially an image segmentation based on geometric and statistical features, which combines the segmentation and recognition of objects of interest in the picture. Its accuracy and real-time performance are an important indicator of the system. In recent years, target Detection of face recognition, unmanned driving and other fields have been widely used. However, during the detection process, the acquisition equipment will be affected by factors such as angle, occlusion, light intensity, weather, etc., which will lead to image distortion of the target to be recognized, that is, the change of feature distribution, which adds new challenges to target detection.

It divides the detection problem into two stages. First, the region proposals (Region Proposals) are generated, then they are classified, and the positions are refined. The typical representative of this type of algorithm is the R-CNN algorithm, such as R-CNN, Fast R-CNN, Faster R-CNN, etc.

The classic two-stage target detection model Faster R-CNN(Ren et al. 2015) can be divided into four modules:

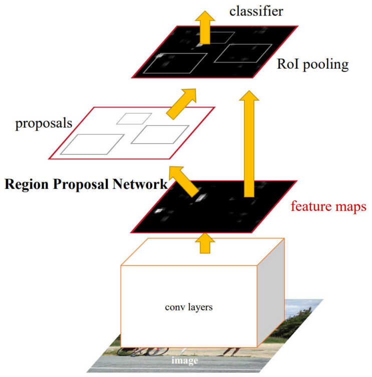

Figure 1: Basic structure of Faster R-CNN(Ren et al. 2015)

The main performance indicators of the target detection model are accuracy and speed. For accuracy, target detection must consider both the classification accuracy of each target and the accuracy of the bounding box positioning. Accuracy and speed are often inversely proportional, how to balance the two is an important direction of target detection algorithm research.

Theory of Domain Generalization

Basic Theory of Domain Generalization The problem of domain generalization (Domain Generalization, DG) research is to learn a model with strong generalization ability from several data sets (domains) with different data distributions, so as to achieve better results on the unknown test set.

The biggest difference between domain generalization and domain adaptation (Domain Adaptation, DA): domain adaptation in training, both source domain and target domain data can be accessed (only unlabeled target domain data in unsupervised domain adaptation); In the generalization problem, we can only access several source domain data for training, and the test data cannot be accessed. There is no doubt that domain generalization is a more challenging and practical scenario than domain adaptation, and practical applications tend to produce machine learning models that are sufficiently generalized in one training session.

Definition of Domain Generalization We can define the domain generalization problem as shown in 2. The following will briefly introduce the concept of domain and domain generalization.

Let $\mathcal{X}$ denote a non-empty input space, $\mathcal{Y}$ denote an output space, and the domain consists of data sampled from the corresponding dataset distribution. we denote it as

$$\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \sim P_{XY} \quad (1)$$

Where $\mathbf{x} \in \mathcal{X} \subset R^d$, $y \in \mathcal{Y} \subset R$ represents the label, P_{XY} represents the joint distribution of input samples and

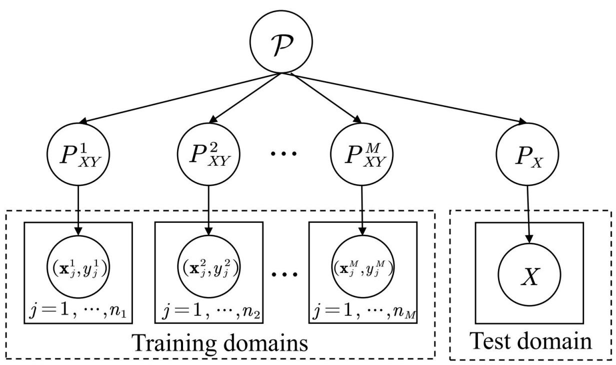

Figure 2: Domain generalization diagram(Wang et al. 2021)

output labels. X and Y represent the corresponding random variables.

In the definition of domain generalization, we assume that given M training source domains

$$\mathcal{S}_{\text{train}} = \left\{\mathcal{S}^i \mid i = 1, \cdots, M\right\} \tag{2}$$

where $\mathcal{S}^i = \left\{\left(\mathbf{x}^i_j, y^i_j\right)\right\}_{j=1}^{n_i}$ represents the ith domain, and the domains are subject to different distributions, that is, $P^i_{XY} \neq P^j_{XY}, 1 \leq i \neq j \leq M$. The goal of domain generalization is to learn a generalized prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ such that it is in the target domain $\mathcal{S}_{\text{test}}$ ($P^{\text{test}}_{XY} \neq P^i_{XY}$ for $i \in \{1, \cdots, M\}$) has the smallest prediction error, namely:

$$\min_h E_{(\mathbf{x},y)\in\mathcal{S}_{\text{test}}}\left[\ell(h(\mathbf{x}), y)\right] \tag{3}$$

Where E and $\ell(\cdot,\cdot)$ are both loss functions.

Theory of Graph Convolutional Network

Basic Theory of Graph Neural Network In computer science, the graph data structure is composed of two parts: node (Vertice) and edge (Edge). A graph G is described by a set of vertices V and a set of edges E contained in the graph. The direction dependency between nodes determines whether the edge is directed or undirected. Nodes can also be called vertices. In this paper, the two are equivalent.

Graph Neural Network (GNN) is a neural network built on the basis of graph structure. A common application scenario of GNN is node classification. Each node in the graph we construct is associated with a label. There are unlabeled nodes in the graph. We hope to classify unlabeled nodes through nodes with known labels.

Any node v of the graph is represented by its feature $\mathbf{x}_v$, and the label $\mathbf{t}_v$ that has been annotated is associated with it, through a given part The marked graph G uses the marked nodes to predict the unmarked node labels, and obtains the state information $\mathbf{h}_v$ of each node through training. In the continuous aggregation operation $\mathbf{h}_v$ also contains information about neighboring nodes.

$$\mathbf{h}_v = f\left(\mathbf{x}_v, \mathbf{x}_{co[v]}, \mathbf{h}_{ne[v]}, \mathbf{x}_{ne[v]}\right) \tag{4}$$

$\mathbf{x}_{co[v]}$ represents the characteristics of the edge connected to the vertex v, $\mathbf{h}_{ne[v]}$ represents the neighbor node of the vertex v State information, $\mathbf{x}_{ne[v]}$ represents the neighbor node characteristics of vertex v. f is the transfer function that projects the input to the d dimensional space. For each aggregation and update operation of $\mathbf{h}_v$, rewrite the above equation for iterative update as follows.

$$\mathbf{H}^{t+1} = F\left(\mathbf{H}^t, \mathbf{X}\right) \tag{5}$$

$\mathbf{H}$ and $\mathbf{X}$ denote all connections of $\mathbf{h}$ and $\mathbf{x}$ respectively, by putting state $\mathbf{h}_v$ And the feature $\mathbf{x}_v$ is passed to the output function g to calculate the output $\mathbf{o}_v$.

$$\mathbf{o}_v = g\left(\mathbf{h}_v, \mathbf{x}_v\right) \tag{6}$$

Both f and g here can be interpreted as a fully connected feedforward neural network.

Algorithm of Graph Convolutional Network Graph Convolutional Network (Kipf and Welling 2016) (Graph Convolutional Network, GCN) was proposed in 2017. The operation of convolution on the graph provides a new idea for processing graph-structured data, and provides a combination of digital images and graph data in deep learning. Widen the road.

Graph convolutional networks can be divided into spectral convolutions and spatial convolutions(Niepert, Ahmed, and Kutzkov 2016). Spectral convolution moves the filter and graph data in the network to the Fourier domain at the same time for processing. The convolution in the spatial domain is more intuitive, directly connecting the nodes of the graph in the spatial domain to form a hierarchical structure and perform convolution.

For graph G, its Laplacian matrix is defined as

$$L = D - A \tag{7}$$

Where L is the Laplacian matrix, D is the degree matrix, and the elements on the diagonal are the degree of the vertex, that is, the number of elements linked by the element, A It is an adjacency matrix (Adjacency matrix), that is, it represents the adjacency relationship between any two vertices, and the adjacency is 1, and the non-adjacency is 0.

The connection relationship of graph $G = (V, E)$ is reflected in the adjacency matrix A and degree matrix D of the graph. Usually, we need to normalize the Laplacian matrix to get a symmetric normalized matrix

$$L^{sys} = D^{-1/2} L D^{-1/2} \tag{8}$$

Our analysis of the graph can be regarded as the analysis of its Laplacian matrix.

A Domain Generalized Object Detection Model Based on Graph Convolutional Network

Task Description

This paper follows the setting of the common domain generalization task (Li et al. 2018). where $\mathcal{X}$ and $\mathcal{Y}$ denote the input space and label space respectively, assuming that there are K visible on the joint space $\mathcal{X} \times \mathcal{Y}$ Source domain $\mathcal{D} = \{D_1, D_2, \cdots, D_k\}$. This work focuses on common classification tasks, therefore, $\mathcal{Y}$ is the domain-shared set of discrete numbers $\mathcal{Y} = \{1, 2, \cdots, C\}$, where C Indicates the number of categories contained in the dataset. As for the input space $\mathcal{X}$, the data sample is from the dataset $D_k = \{(x_n, y_n)\}_{n=1}^{N_k}$, where N_k is the total number of labeled samples in the source domain D_k. The goal of the domain generalization task is to use multiple visible source domains $\mathcal{D}$ to learn a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ that generalizes well to any novel target domain $\mathcal{D}_{te}$, where the multiple source and target domains cover the same set of categories but have different distributions of statistics. During model learning, the domain generalization task assumes that the target domain does not have any available information, and the model focuses on learning a domain-invariant but class-dependent latent feature space based on all source domain data.

Motivation

The domain generalization task focuses on learning a model with strong generalization ability and good performance in completely unknown data distribution applications through several data sets with different data distribution characteristics. In the field of computer vision and natural language processing, because the model training data can only access several source domain data with the same task and different distributions, but the test data cannot be accessed, so most of the current existing work focuses on designing models suitable for domains. training methods, and they are often evaluated on image classification datasets.

The object detection task is to find out all the objects artificially set in the digital image and determine their category and location, and its focus is on:

1. Classification problem: which category does a picture or image in a region belong to.
2. positioning problem: the target may appear anywhere in the image.
3. size issue: objects come in various sizes.
4. shape problem: objects may have various shapes.

Since various objects have different appearances and are interfered by factors such as light and weather during imaging, and the sensor models of the acquisition equipment are different, the images used as training data are often not satisfied with the same distribution, resulting in the same distribution of data. The model trained on the above performs poorly when applied to another data with a different distribution. This phenomenon is especially obvious in the field of object recognition.

The popular standard data sets in the current domain generalization problem mainly include Office-Caltech, PACS, VLCS, etc., without exception, are data sets for classification tasks, although domain generalization is widely used in semantic segmentation, medical image processing and other fields , but most of the current operations are training after simulated domain shift through data augmentation.

At present, there is no dedicated data set for the generalization task of target recognition domain. Therefore, in order to achieve better target detection results in source domain data with different distributions, this paper will build a self-built data set and collect data with different distributions. The characteristic images are self-labeled to meet the experimental needs and provide training data convenience for researchers in related fields.

The graph neural network models unstructured data into structured graphs, and uses these graphs to generate higher-level, more generalized models. Its brain-like information processing method does not pay too much attention to the uniqueness of a certain data set. Instead, it focuses on inter-class features with stronger universality, which is conducive to enhancing the generalization ability of the model and has the characteristics of learning to learn.

Based on the above, the goal of this paper is to collect and label a data set suitable for the generalization of the target recognition domain, and to provide a target detection algorithm based on graph convolutional networks that has good performance on source domain data with different distributions.

Proposed Modules

The algorithm given in this paper is verified on the Faster-Rcnn model with VGG16 as the backbone network. Based on the challenges faced by the above-mentioned target detection domain generalization task and the advantages of the graph convolutional network, this paper focuses on three aspects: pixel level, instance level, and domain level. Each level constrains the inherent semantics of digital images, and fully explores the correlation between categories and beyond categories in the process of target detection. The overall structure of the network is shown in 3, and the FasterRcnn network structure is not fully displayed.

Pixel-level image convolutional network: extract the high-level feature map in the backbone network, downsample to an appropriate size (20*20), use the relative distance between pixels as the weight, and weight the object area in the image to obtain the target The pixel distance of the target. After that, the adjacency matrix of the graph convolutional network is constructed based on the cosine similarity between the pixels of the feature map, and the pixel distance is constrained so that the adjacency matrix acts on the pixels containing the target, and the fused features are input into the fully connected layer to obtain pixel-level The loss is used to guide the classification.

In this layer of graph convolutional network, we agree on the functions and their corresponding functions as follows:

B is used to judge whether the pixel in the current feature map is the marked real object frame

D is used to calculate the relative distance of each pixel in the feature map

S is used to calculate cosine similarity

E Cross entropy loss function

C classifier

N regularization function

The pixel-level image convolutional network first constructs the relationship between the middle and high-level feature map $\mathbf{x_p}$ and the pixels of the real label frame to form a mask related to the pixel distance and the real target:

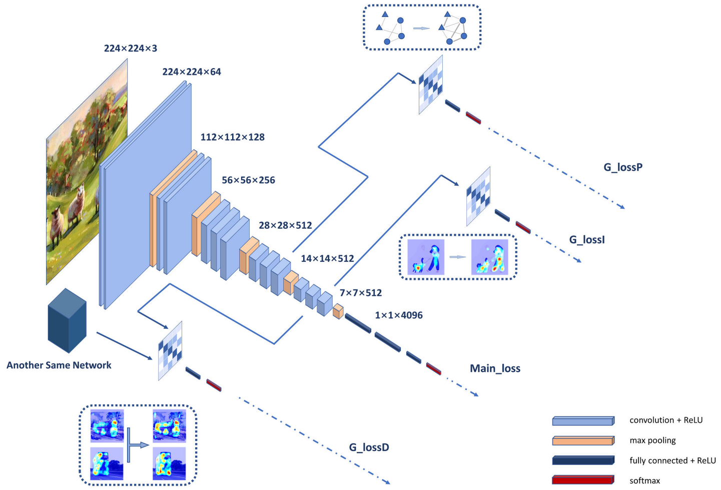

Figure 3: The overall structure of the model

$$Mask_{pixel} = B(\mathbf{x_p}) \odot D(\mathbf{x_p})$$

Then use the cosine similarity to explore the internal correlation of the feature map pixels, and use the above mask as the similarity weight to construct the adjacency matrix between pixels:

$$Matrix_{simi} = N(Mask_{pixel} \odot S(\mathbf{x_p}, \mathbf{x_p}^T))$$

The pixel-level graph convolutional network module can be regarded as GCN_p, and the feature map is assigned to the original feature after the feature fusion between each pixel through the adjacency matrix:

$$\mathbf{x_p} = GCN_p(\mathbf{x_p}, Matrix_{simi})$$

The fused features and original labels compute the pixmap convolutional cross-entropy loss for backpropagation:

$$Loss_{pixel} = E(C(\mathbf{x_p}), \mathbf{label})$$

Instance-level graph convolutional network: In the Faster R-cnn model, the pooled features are sorted by confidence, select instances with higher confidence, and strengthen the semantic constraints of different parts of the same instance by constructing the internal adjacency matrix of the instance. That is, to enhance the internal correlation of the category, the instance after feature fusion is input into the fully connected layer, and the instance-level loss is obtained to guide the classification.

The agreed functions and their functions in this hierarchical graph convolutional network are the same as above. For this layered network, the newly added functions and their functions are as follows:

Q Calculate the similarity and difference matrix of labels between instances

Instance-level masks associated with classes are available:

$$Mask_{instance} = Q(\mathbf{label}, \mathbf{label^T})$$

For instance-level features $\mathbf{x_i}$, we can construct such an adjacency matrix to strengthen the relationship between different types of features of the same category instance:

$$Matrix_{simi} = N(Mask_{instance} \odot (1 - S(\mathbf{x_i}, \mathbf{x_i}^T)))$$

The instance-level graph convolutional network module can be regarded as GCN_i, which fuses the instance features with high confidence through the adjacency matrix and assigns them to the original features:

$$\mathbf{x_i} = GCN_i(\mathbf{x_i}, Matrix_{simi})$$

The fused instance-level features and original labels compute the instance graph convolutional cross-entropy loss for backpropagation:

$$Loss_{instance} = E(C(\mathbf{x_i}), \mathbf{label})$$

Domain-level graph convolutional network: The above two modules are commonly used in most target detection tasks. For the datasets studied in this paper, there are obvious domain shifts. In order to make full use of domain labels, the instance features from the two domains are in the Under the guidance of the label, select the characteristics of each category that can best represent the instance and domain, and construct an adjacency matrix under cross-domain semantic constraints. The fused features from the two domains are input into the fully connected layer to obtain domain-level losses. Used to guide classification.

The agreed functions and their functions in this hierarchical graph convolutional network are the same as above. For this layered network, the newly added functions and their functions are as follows:

W is used to calculate the single-domain instance-level adjacency matrix weight

[] represents the matrix splicing operation

For cross-domain features $\mathbf{x_{d1}}$ and $\mathbf{x_{d2}}$, and their corresponding labels $\mathbf{label_{d1}}$ and $\mathbf{label_{d2}}$, first get its single-domain adjacency matrix weight:

$$Weight_{d1} = W(\mathbf{x_{d1}}, \mathbf{label_{d1}})$$
$$Weight_{d2} = W(\mathbf{x_{d2}}, \mathbf{label_{d2}})$$

In order to form a cross-domain graph convolutional network, we perform matrix splicing operations on cross-domain features, cross-domain labels, and cross-domain weights:

$$\mathbf{x_d} = [\mathbf{x_{d1}}, \mathbf{x_{d2}}]$$
$$\mathbf{label_d} = [\mathbf{label_{d1}}, \mathbf{label_{d2}}]$$
$$Weight_d = [Weight_{d1}, Weight_{d1}]$$

At this point an adjacency matrix can be constructed:

$$Matrix_{simi} = N((1 - Weight_d) \odot S(\mathbf{x_d}, \mathbf{x_d}^T))$$

The cross-domain graph convolutional network module can be regarded as GCN_d, which is assigned to the original feature after cross-domain feature fusion through the adjacency matrix:

$$\mathbf{x_d} = GCN_d(\mathbf{x_d}, Matrix_{simi})$$

The fused cross-domain features and the original cross-domain labels calculate the instance graph convolution cross-entropy loss for backpropagation:

$$Loss_{domain} = E(C(\mathbf{x_d}), \mathbf{label_d})$$

For the above three image convolutional network loss functions, add coefficients to control the degree of gradient descent, so that the pixel image convolutional network loss $Loss_{pixel}$ coefficient is α_P, similarly, the example map The convolutional network loss $Loss_{instance}$ coefficient is α_I, and the cross-domain graph convolutional network loss $Loss_{domain}$ coefficient is α_D. The total loss coefficient is set to β_G, the loss in the Faster R-cnn model is defined as $Loss_{main}$, and the total loss function of the model can be obtained as follows:

$$\begin{aligned} Loss = Loss_{main} + \beta_G(&\alpha_P Loss_{pixel} \\ &+\alpha_I Loss_{instance} \\ &+\alpha_D Loss_{domain}) \end{aligned} \tag{9}$$

In addition, in view of the above-mentioned differences in the amount of module information and task complexity, the pixel-level graph convolutional network GCN_p adopts a smaller-scale graph convolutional network, and the number of hidden layer nodes is set to 256, 128; instance-level GCN_i and domain-level graph convolutional network GCN_d are larger in scale, and the number of hidden layer nodes is 2048 and 1024. At the same time, in order to prevent over-fitting, 10% of the nodes in the network will be randomly set as inactive nodes.

Experimental Validation

In this chapter, experimental verification of the model introduced above will be carried out. This chapter firstly introduces the design of the data set for cross-domain target recognition in detail, and then provides the experimental results on this data set to verify the algorithm proposed in Chapter 3.

Self-made Domain Generalized Datasets of Object Detection

In the domain generalization research, a large number of scholars focus on improving the generalization ability of models in natural image classification tasks, and put forward a series of solutions and data sets. Common natural image datasets used for domain generalization of classification problems include handwritten character dataset MNIST and target recognition benchmark dataset VLCS(Fang, Xu, and Rockmore 2013). No data set for target detection domain generalization task has yet appeared. Therefore, this paper collected and labeled pictures from three domains as experimental data set, combined with VOC2007 data set, a total of four data sets with different parts for experimental verification and exploration.

The dataset we collected and labeled is used for the study of generalization of target recognition domain, such as **??**, which contains four domains with very serious distribution differences – photos (VOC2007), sketches (sketch22), watercolors (watercolor22), and cartoon (clipart22). There are 20 categories of samples: motorbike, bird, pottedplant, boat, car, person, cow, chair, bus, bicycle, diningtable, tvmonitor, bottle, aeroplane, sheep, dog, horse, cat, sofa, train.

In order to ensure the accuracy, effectiveness and stability of the domain generalization target detection algorithm verified on this dataset as much as possible, it is necessary to label as many and accurate images as possible, as shown by 1. In this paper, a total of 4050 pictures in sketch22, watercolor2021 and clipart22 were marked with 9720 detection boxes, each picture contained 2.4 objects of interest.

In object detection, we want to classify high-dimensional data more accurately, so the model we build should be able to optimize the feature spatial distribution of data with different distributions, that is, reduce the spacing between the same type and expand the spacing between different types.

Experimental Validation on Cross Domain Datasets

Implementation Details There are obvious distribution differences among the domains of the self-made data set, which is very challenging. The model is based on the pre-trained VGG16 on ImageNet as the backbone network. In the experiment, batch size of the training set $bs = 2$, initial learning rate $lr = 1e - 3$, number of training rounds $epochs = 7$, and $iters$= the number of pictures/batch size in

Table 1: Number of tags and images per field

	VOC2007	sketch22	watercolor22	clipart22
motorbike	759	31	24	119
bird	1175	138	181	253
pottedplant	1217	113	176	468
boat	791	117	121	151
car	3185	82	180	276
person	10674	447	477	2030
cow	685	99	99	145
chair	2806	242	144	494
bus	526	31	43	122
bicycle	807	48	57	119
diningtable	609	50	19	137
tvmonitor	728	48	36	142
bottle	1291	77	70	241
aeroplane	642	67	48	160
sheep	664	81	88	151
dog	1068	54	71	136
horse	801	121	141	137
cat	759	51	24	142
sofa	821	58	31	107
train	630	72	65	65
Picture number	9963	1000	1050	2000

each training round. In order to promote the convergence of the model, the learning rate was adjusted to $lr = 1e - 4$after 5 rounds of training.

For the multiple loss functions used in the classification of auxiliary networks in this paper, the degree of gradient descent is controlled by coefficient, and the coefficients $\alpha_P = 0.05$, $\alpha_I = 0.1$, $\alpha_D = 0.1$, The convolution loss coefficient of the total graph is set to $\beta_G = 0.2$.

Experimental Contents The problem to be solved by the model proposed in this paper is target recognition domain generalization task. In order to further prove the universality of the method proposed in this paper, the experimental performance of this method is verified on the above data set. In order to compare the performance of the methods, the experiment follows the experimental setting of "Leave one method" (Li et al. 2018), that is, assuming that there are Ndifferent domains in the data set, $n - 1$domains are selected as the source domain, and the remaining one domain is the target domain for testing. In addition, in order to better verify the influence of the proposed algorithm on the model, some ablation experiments are also added in this paper for further exploration and analysis.

The experimental code language of this chapter is Python, and the convolutional neural network is implemented based on Pytorch. The code runs on the following hardware configurations: Intel(R) Xeon(R) CPU, 2.30GHz main frequency, 32GB system memory, NVIDIA P100 video card, 16GB video memory.

In target detection tasks, unilateral evaluation indexes are commonly used as follows:

(Precision): TP/(TP + FP)

(Recall): TP/(TP + FN)

PR curve: Precision-Recall curve

AP: Area under PR curve

In this experiment, mAP(mean Average Precision), a common comprehensive performance index used in target detection, was used to evaluate the performance, that is, the average AP of each category.

In this experiment, the benchmark algorithm DeepAll is compared with the algorithm proposed in this paper. 2 summarizes the target recognition results on the self-made data set. The experimental results of the benchmark algorithm come from the same parameter Settings. The statistical results prove that the performance of the proposed algorithm is improved in multi-domain target detection tasks. Compared with the DeepAll algorithm, the proposed method improves the mAP by 3.92%. In this method, various kinds of features in the training process are modeled through the graph relational network, thus improving the ability of the model to be extended to unknown domains.

It is worth mentioning that this paper has better generalization effect on more abstract target domains. For example, in the generalization task with sketch as the target domain with the maximum domain offset, mAP of the method in this paper is 4.97% higher than the benchmark algorithm, and in the task with clipart as the target domain, it is even 7.41% higher. It fully embodies the excellent generalization performance of the model.

mAP index numerically illustrates the good performance of the algorithm in this paper. In order to demonstrate the great improvement of the algorithm in the visual experience of human eyes, the gradient-based method (Selvaraju et al. 2017) is used to draw the feature map. The warmer the heat map is, the more the network pays attention to this feature. As shown in 4, 5, 6, 7, When detecting different objects, the network sometimes pays too much attention to the partial features of a certain kind of object rather than the overall semantic information of the object, which is easy to affect the classification of the network and the positioning of the detection frame under the circumstances of occlusion and illumination change. After modeling the image convolutional network of pixel level, instance level, domain level and object, it can be clearly seen that the coverage of thermal map becomes larger and the focus of feature points becomes more. The algorithm in this paper can guide the network to observe the universality characteristics of a certain type of object rather than focusing on the local area, so that the network has better generalization performance.

In addition, the proposed algorithm also made some achievements in improving the recall ratio of the network. As shown in 8, the benchmark algorithm DeepAll could not detect obvious targets in the image, such as sofa and bus occupying a large area of the image. The proposed algorithm could correctly detect and improve the recall ratio. When there are occlusions between objects, the proposed algorithm can detect the occluded objects to a certain extent, which again demonstrates the good generalization performance of the model.

Ablation Experiment The key components of the method in this paper include the graph convolutional network at three levels and its corresponding losses. In order to better understand the effect of each component of the model, the

Table 2: Performance comparison between the textual model and the DeepAll method

2*Task	Source Domain	S,W,V	C,W,V	C,S,V	C,S,W	2*mAP
2*	Target Domain	C	S	W	V	2*
2*Model	DeepAll	50.90	66.42	63.65	53.96	58.73
2*	Ours	54.67	69.72	65.62	54.12	61.03

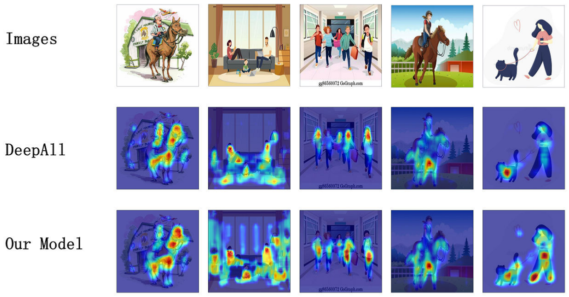

Figure 4: Clipart

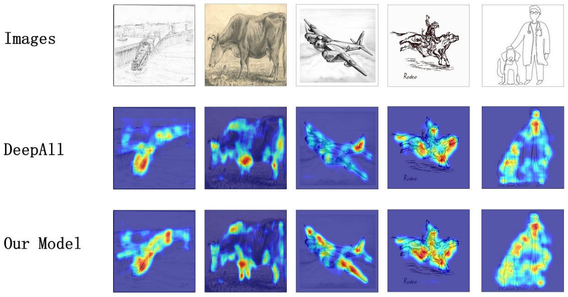

Figure 5: Sketch

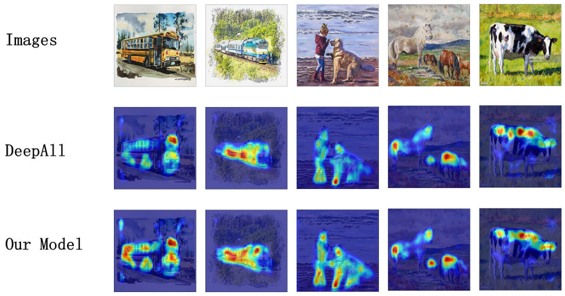

Figure 6: Watercolor

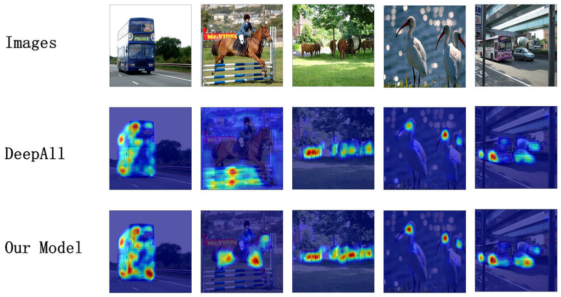

Figure 7: VOC

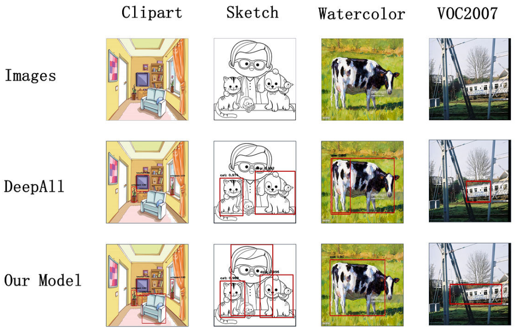

Figure 8: Model detection performance comparison

control variable method is used to conduct ablation research on each key component.

3 shows the statistical results obtained by removing any level of the three-level graph convolution network respectively. It can be obviously observed that mAP index is significantly lower than that of the three modules when removing any module, especially after removing G_{pixel}, mAP decreases by 1.33%. After removing $G_{instance}$, it decreases by 1.25%. After the above two modules are removed respectively, the generalization task with sketch as the target domain is most affected. It can be inferred that semantic constraints at pixel level and instance level are more important in data sets with large domain offset.

In addition, the total loss coefficient of the graph convolutional network is a super parameter, which needs to be set artificially. Experiments are also conducted in this paper to prove that the model performs best when $\beta_G = 2.0$. 4 is the experimental result of using control variable method.

Table 3: Study on ablation of each Module of the model

2*G_{pixel}	2*$G_{instance}$	2*G_{domain}	Source Domain	S,W,V	C,W,V	C,S,V	C,S,W	2*mAP
2*	2*	2*	Target Domain	C	S	W	V	2*
				54.36	67.89	64.91	53.70	60.22
				54.02	68.04	65.15	53.87	60.27
				54.14	69.28	66.11	54.33	60.97

Table 4: Study on the Total Coefficients $\beta_G$of Graph Convolutional Networks

2*Task	Source Domain	S,W,V	C,W,V	C,S,V	C,S,W	2*mAP
2*	Target Domain	C	S	W	V	2*
2*	1.0	54.52	68.29	66.03	54.15	60.75
2*β_G	2.0	54.67	69.72	65.62	54.12	61.03
2*	3.0	54.42	68.34	65.76	53.29	60.45
2*	4.0	54.31	68.33	65.39	52.89	60.23
2*	5.0	53.52	68.74	66.17	52.30	60.18

Conclusion and Future Works

Conclusion

Traditional deep learning requires a large number of manually marked data, and needs to meet the premise that training and testing data follow independent and same distribution. In practical application, marked data is not easy to obtain. In order to reduce the dependence on marked data, the concept of domain generalization is proposed, hoping to make full use of multiple visible source domain data to make the model still perform well in unknown domain. In the past, domain generalization tasks were mostly carried out on the classification tasks, and there was a certain distance between the tasks in contact with the actual production and life, such as the target detection task, and the target detection task was more likely to be blocked by light and other fields to produce domain deviation, thus resulting in poor model performance.

To solve the above problems, this paper proposes a domain generalization model for target detection tasks, so as to improve the generalization ability of the target detection model. The model uses the graph convolutional network to conduct semantic constraints on the semantic information at the pixel level, instance level and domain level, and strengthens the connection from the same category features, so that the model focuses on the general features of the category itself, rather than the local features, and enrichis the generalization experience.

The method proposed in this paper effectively optimizes the feature space, further explores the information contained in the digital image itself, uses the graph convolution method to spontaneously converge, fuse and update features, and carries out semantic constraints. In order to verify the performance of the proposed method, the corresponding experimental verification and comparative analysis are carried out on the self-built data set, and the statistical results prove the effectiveness of the proposed method.

Future Works

This paper proposes a domain generalization model for target detection tasks, learns domain-independent features by means of graph convolution, accumulates more generalization experience, and achieves superior performance in the face of more abstract target domains. However, there are still some problems to be solved in this model, which can be improved from the following aspects in the future:

First, the research focus of this paper is the task of generalization from multi-source domain to single domain, which reduces the dependence of deep learning on labeled samples. Currently, in order to further reduce the cost of labeling, new tasks such as generalization from single source domain to multi-target domain and generalization from multi-source domain to multi-target domain are proposed, which have not been covered in this paper. How to do well the generalization of multi-kind data sets is a meaningful research direction.

Second, in the three-level graph convolutional network constructed in this paper, pixel level and instance level networks play an important role in the model, while the role of cross-domain graph convolutional network has not been fully explored. How to make good use of domain labels to reasonably model and constrain the semantic information representing each domain may bring greater performance improvement.

References

Arruda, V. F.; Paixão, T. M.; Berriel, R. F.; De Souza, A. F.; Badue, C.; Sebe, N.; and Oliveira-Santos, T. 2019. Cross-domain car detection using unsupervised image-to-image translation: From day to night. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Cai, Q.; Pan, Y.; Ngo, C.-W.; Tian, X.; Duan, L.; and Yao, T. 2019. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11457–11466.

Chen, C.; Zheng, Z.; Ding, X.; Huang, Y.; and Dou, Q. 2020. Harmonizing transferability and discriminability for adapting object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8869–8878.

Chen, C.; Zheng, Z.; Huang, Y.; Ding, X.; and Yu, Y. 2021. I3net: Implicit instance-invariant network for adapting one-stage object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12576–12585.

Chen, Y.; Li, W.; Sakaridis, C.; Dai, D.; and Van Gool, L. 2018. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3339–3348.

Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE International Conference on Computer Vision*, 1657–1664.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1): 2096–2030.

Huang, P.; Han, J.; Cheng, D.; and Zhang, D. 2022. Robust Region Feature Synthesizer for Zero-Shot Object Detection. *arXiv preprint arXiv:2201.00103*.

Khodabandeh, M.; Vahdat, A.; Ranjbar, M.; and Macready, W. G. 2019. A robust learning approach to domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 480–490.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. M. 2018. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Lin, C.-T. 2019. Cross domain adaptation for on-road object detection using multimodal structure-consistent image-to-image translation. In *2019 IEEE international conference on image processing (ICIP)*, 3029–3030. IEEE.

Liu, B. 2017. Lifelong machine learning: a paradigm for continuous learning. *Frontiers of Computer Science*, 11(3): 359–361.

Niepert, M.; Ahmed, M.; and Kutzkov, K. 2016. Learning convolutional neural networks for graphs. In *International conference on machine learning*, 2014–2023. PMLR.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shuman, D. I.; Narang, S. K.; Frossard, P.; Ortega, A.; and Vandergheynst, P. 2013. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3): 83–98.

Wang, J.; Lan, C.; Liu, C.; Ouyang, Y.; Zeng, W.; and Qin, T. 2021. Generalizing to unseen domains: A survey on domain generalization. *arXiv preprint arXiv:2103.03097*.

Zhu, X.; Pang, J.; Yang, C.; Shi, J.; and Lin, D. 2019. Adapting object detectors via selective cross-domain alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 687–696.