Disentangle and Remerge: Interventional Knowledge Distillation for Few-Shot Object Detection from A Conditional Causal Perspective

Jiangmeng Li[12*], Yanan Zhang[12*], Wenwen Qiang[12†], Lingyu Si[12], Chengbo Jiao[3], Xiaohui Hu[2], Changwen Zheng[2], Fuchun Sun[4]

[1]University of Chinese Academy of Sciences
[2]Institute of Software Chinese Academy of Sciences
[3]University of Electronic Science and Technology of China
[4]Tsinghua University
{jiangmeng2019, yanan2018, qiangwenwen, lingyu, hxh, changwen}@iscas.ac.cn,
chengbojiao@hotmail.com, fcsun@mail.tsinghua.edu.cn

Abstract

Few-shot learning models learn representations with limited human annotations, and such a learning paradigm demonstrates practicability in various tasks, e.g., image classification, object detection, etc. However, few-shot object detection methods suffer from an intrinsic defect that the limited training data makes the model cannot sufficiently explore semantic information. To tackle this, we introduce knowledge distillation to the few-shot object detection learning paradigm. We further run a motivating experiment, which demonstrates that in the process of knowledge distillation, the empirical error of the teacher model degenerates the prediction performance of the few-shot object detection model as the student. To understand the reasons behind this phenomenon, we revisit the learning paradigm of knowledge distillation on the few-shot object detection task from the causal theoretic standpoint, and accordingly, develop a Structural Causal Model. Following the theoretical guidance, we propose a backdoor adjustment-based knowledge distillation method for the few-shot object detection task, namely ***D**isentangle and **R**emerge* (D&R), to perform conditional causal intervention toward the corresponding Structural Causal Model. Empirically, the experiments on benchmarks demonstrate that D&R can yield significant performance boosts in few-shot object detection. Code is available at https://github.com/ZYN-1101/DandR.git.

1 Introduction

Learning robust and generic representations with limited labels is a long-standing topic in machine learning. Few-shot learning, an innovative representation learning paradigm, is practicable in various tasks, e.g., image classification (Finn, Abbeel, and Levine 2017; Vinyals et al. 2016; Snell, Swersky, and Zemel 2017; Sung et al. 2018; Chen et al. 2019a), object detection (Yan et al. 2019; Kang et al. 2019; Wang et al. 2020; Qiao et al. 2021; Zhu et al. 2021), etc.

In general, researchers explore approaches to tackle object detection problems under the setting of few-shot learning in two promising directions: 1) the meta-based methods train the model via a huge amount of few-shot detection tasks sampled from the base classes; 2) the fine-tune-based methods aim to transfer the knowledge from base classes to novel classes. Due to the intrinsic limitation of few-shot learning, it is challenging for the model to sufficiently explore semantic information from the input data. Therefore, we introduce knowledge distillation (Hinton, Vinyals, and Dean 2015) to improve the ability of few-shot object detection (FSOD) models to acquire semantic information by learning from large-scale pre-trained models, such as CLIP (Radford et al. 2021). According to the knowledge distillation learning paradigm, minimizing distillation loss entails aligning the distribution of classification *logits* generated by the *teacher* model and *student* model, and thus both the "correct" and "incorrect" knowledge of the teacher model is learned by the student model in the learned feature space. From the foundational principle of knowledge distillation, the distillation loss can be considered as an auxiliary loss to improve the performance of the main model, and further, in the case that the teacher model has a stronger semantic capturing ability than the student model, the knowledge dis-

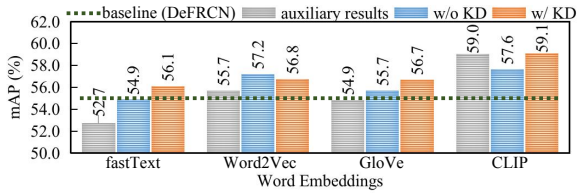

Figure 1: Comparisons of FSOD models enhanced by auxiliary approaches. Besides the main FSOD task, we introduce an auxiliary task, which encodes the *categories* by the auxiliary approaches and then trains the feature extractor by backpropagating the cross-entropy loss based on the embedded categories and visual features learned by the feature extractor. The inference results are achieved by using schemes: 1) *auxiliary results* are the auxiliary outputs; 2) the results of *w/o KD* are the main outputs; 3) for *w/ KD*, we introduce the knowledge distillation in fine-tuning, and the results are the main outputs. Refer to Figure 4 for architecture details.

*These authors contributed equally.
†Corresponding author.

tillation usually has a considerable promotion on the main model. Yet, observations from the motivating experiments in Figure 1, on the other hand, contradict this.

Specifically, we run the motivating experiments on the VOC dataset (Everingham et al. 2010) with Novel Set 1 by using benchmark methods, including fastText (Pennington, Socher, and Manning 2014), Word2Vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), and CLIP (Radford et al. 2021). The results are demonstrated in Figure 1. We observe that the variant of w/ KD generally outperforms the compared variants. The auxiliary results, as the control group, are the lowest on most tasks. According to our statement, CLIP, as a large-scale vision-language model, can better improve the model's ability to capture semantic information by using knowledge distillation. However, there exists a counterintuitive phenomenon: for Word2Vec, the w/ KD variant underperforms the w/o KD variant. A plausible explanation is that in the process of knowledge distillation, the FSOD model, as the student model, not only learns the knowledge of the teacher model for the acquisition of open-set semantic information, but also the empirical error of the teacher model degenerates the student model's prediction of the target labels. The teacher's quality severely affects the performance of the student, and several specific teachers may not improve the performance of the student, but instead interfere with the student's predictions on downstream tasks. This is in accordance with the observation of Figure 1. Therefore, such a reason may degenerate the performance of all knowledge distillation-based models, including CLIP-based models.

To tackle this issue, we revisit the learning paradigm of knowledge distillation on the FSOD task from the causal theoretic standpoint. Accordingly, we develop a Structural Causal Model (SCM) (Pearl 2009; Glymour, Pearl, and Jewell 2016) [1] to describe the causal relationships between the corresponding variables in this paper. As demonstrated in Figure 2, the proposed SCM focuses on exploring the causal graph of knowledge distillation-related variables, including the candidate image data, whole open-set semantic knowledge of the teacher model, classification knowledge for downstream tasks, general discriminant knowledge for distinguishing foreground and background objects, and target label. When analyzing the SCM, we discover that it is an exception to current causal inference approaches and that the existing standard definition of the backdoor criterion has limitations, to a certain extent. Inspired by recent works (Van der Zander, Liskiewicz, and Textor 2014; Perkovic et al. 2018; Correa and Bareinboim 2017), we propose to expand the backdoor criterion's application boundary on the conditional intervention cases without using extra symbols.

For the detailed methodology, guided by the proposed SCM, we *disentangle* the knowledge distillation objective into four terms. By analyzing the impact of such terms against the SCM, we determine that a specific term can be treated as a *confounder*, which leads the student model to learn the *exceptional* correlation relationship between the classification knowledge and general discriminant knowledge for distinguishing foreground and background objects of the teacher model during knowledge distillation. This is the pivotal reason behind the explanation of the observation in Figure 1, i.e., interfering with the student's predictions. Then, to eliminate the negative impact of the confounder and execute conditional causal intervention toward the proposed SCM, we remove the confounder term and *remerge* the remaining terms as the new knowledge distillation objective. We name the proposed backdoor adjustment-based approach ***D**isentangle and **R**emerge* (D&R). Our experiments on multiple benchmark datasets demonstrate that D&R can improve the performance of the state-of-the-art FSOD approaches. The sufficient ablation study further proves the effectiveness of the proposed method. Our major contributions are four-fold:

- We introduce the knowledge distillation to improve the ability of FSOD models to acquire semantic information by learning from large-scale pre-trained models.
- We observe a paradox that adopting different teacher models, knowledge distillation may both promote and interfere with the prediction of the student model.
- To understand the causal effects of the knowledge distillation learning paradigm, we establish the SCM. We propose to expand the backdoor criterion's application boundary on the conditional intervention cases without using extra symbols.
- Guided by the planned SCM, we propose a new method, called Disentangle and Remerge (D&R), by implementing knowledge distillation with backdoor adjustment. Empirical evaluations demonstrate the superiority of D&R over state-of-the-art methods.

2 Related work

2.1 Vision-Language Models

Vision-language models have attracted a lot of attention and shown impressive potential in several areas (Anderson et al. 2018; Antol et al. 2015; Huang et al. 2019; You et al. 2016; Ma et al. 2022). High-quality annotated multi-modal data is often difficult to obtain, so unsupervised learning is preferred nowadays. Typical works (Lu et al. 2019; Tan and Bansal 2019; Chen et al. 2019b; Li et al. 2020) have made tremendous progress in learning universal representations that are easily transferable to downstream tasks via prompting (Jia et al. 2021; Zhang et al. 2020). CLIP (Radford et al. 2021) is one of the most impressive works, which leverages contrastive learning to align the embedding spaces of texts and images using 400 million image-text pairs, and achieves remarkable performance gain in various tasks. We are the first to introduce CLIP into FSOD.

2.2 Few-Shot Object Detection

FSOD aims to build detectors toward limited data scenarios. Meta-based methods (Yan et al. 2019; Kang et al. 2019; Karlinsky et al. 2019) dominate early research. TFA (Wang et al. 2020) outperforms the previous meta-based methods by only fine-tuning the last layer of the detector. After that,

[1] The principal concepts and methodologies are shared by (Pearl 2009) and (Glymour, Pearl, and Jewell 2016).

fine-tune-based methods (Wu et al. 2020; Zhang and Wang 2021) become popular. The most related works to our approach are SRR-FSD (Zhu et al. 2021) and Morphable Detector (MD) (Zhao, Zou, and Wu 2021), which introduce external information to boost the detection of novel classes. Differently, these two methods adopt pure language models to generate semantic embeddings, bringing bias because of the domain gap. Moreover, our method is able to draw on external information more effectively through the distillation loss we proposed.

2.3 Knowledge Distillation

Knowledge distillation is first proposed by the work of (Bucila, Caruana, and Niculescu-Mizil 2006) and (Hinton, Vinyals, and Dean 2015). Generally, knowledge distillation can be divided into three categories: logits-based methods (Hinton, Vinyals, and Dean 2015; Cho and Hariharan 2019; Yang et al. 2019; Zhao et al. 2022), feature-based methods (Romero et al. 2015; Zagoruyko and Komodakis 2017) and relation-based methods (Yim et al. 2017; Tung and Mori 2019). Feature-based methods and relation-based methods achieve preferable performance nowadays. Zhao et al. (2022) which shows competitive results decouples the loss function of the classical logits-based method and provides insights to analyze the key factors of distillation. Guided by the proposed SCM, we disentangle the knowledge distillation objective and remerge them.

2.4 Causal Inference

In the past few years, causal inference (Pearl 2009; Glymour, Pearl, and Jewell 2016) has been widely applied in various fields such as statistics, economics, and computer science. Specifically, in the area of computer vision, it focuses on eliminating spurious correlations through deconfounding (Lopez-Paz et al. 2017; He, Shen, and Cui 2021) and counterfactual inference (Yue et al. 2021; Chang, Adam, and Goldenberg 2021). Deconfounding enables estimating causal effects behind confounders. Wang et al. (2021) introduces a causal attention module (CaaM) to learn causal features with the unsupervised method. CIRL (Lv et al. 2022) builds a SCM to formalize the problem of domain generation and separates the causal factors from the non-causal factors in the input data to learn domain-independent representations. We introduce causal inference in FSOD and build the SCM to understand its learning paradigm when applying knowledge distillation. Guided by the SCM, we propose D&R to boost performance.

3 Problem Formulation

3.1 Knowledge Distillation for Few-Shot Object Detection

Under the intuition that the open-set semantic knowledge can support the object detection task in the few-shot setting, we propose to introduce the knowledge distillation approach in the fine-tuning phase of the FSOD model. In particular, the vanilla knowledge distillation (Hinton, Vinyals, and Dean 2015) can be formulated as:

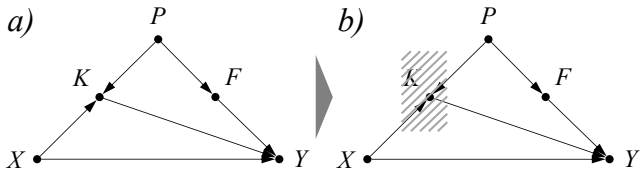

Figure 2: The proposed SCM between candidate image data X, open-set semantic knowledge of the large-scale pre-trained model (e.g., CLIP) P, classification knowledge for specific domains K, general discriminant knowledge for distinguishing foreground and background objects F, and target label Y. a) presents the common causal graph, and b) presents the conditional causal graph, where the causal effect is conditional on K.

$$\mathcal{L}_{\text{KD}} = \text{KL}\left(\mathcal{P}^{\mathcal{T}} \| \mathcal{P}^{\mathcal{S}}\right) = \sum_{i=1}^{N^C} p_i^{\mathcal{T}} log\left(\frac{p_i^{\mathcal{T}}}{p_i^{\mathcal{S}}}\right), \quad (1)$$

where $\mathcal{T}$ and $\mathcal{S}$ denote the teacher model and the student model, respectively. N^C is the number of categories for the FSOD task (including the "background" category). $p_i^{\mathcal{T}}$ and $p_i^{\mathcal{S}}$ denote the classification probabilities generated by the corresponding models using the softmax function. $p_i^{\mathcal{T}}$ and $p_i^{\mathcal{S}}$, as variables, are sampled *i.i.d* from distributions $\mathcal{P}^{\mathcal{T}}$ and $\mathcal{P}^{\mathcal{S}}$, respectively. Note that such a knowledge distillation process is based on the *soft-target* form. We propose to treat the large-scale pre-trained model as the teacher model and the FSOD model as the student model.

3.2 Structural Causal Model

Minimizing the objective formulated by Equation 1 is to make the model learn the object detection and classification knowledge from the large-scale pre-trained model in a distillation manner. From this perspective, minimizing the knowledge distillation objective equals aligning $\mathcal{P}^{\mathcal{T}}$ and $\mathcal{P}^{\mathcal{S}}$ so that both the "correct" and "incorrect" knowledge of the teacher model can be learned by the student model. Then, the SCM implicated in the learning paradigm of knowledge distillation is formalized in Figure 2. The nodes in SCM represent the abstract information variables, e.g., X, and the directed edges represent the (functional) causality, e.g., $X \rightarrow Y$ represents that X is the cause and Y is the effect. In the following, we describe the proposed SCM and the rationale behind its construction in detail at a high level.

$\boldsymbol{X \rightarrow Y \leftarrow K}$. X denotes the candidate image data in a downstream task. Y denotes the corresponding classification label. K denotes the classification knowledge for the specific task. Y is determined by X via two ways: the direct $X \rightarrow Y$ and the mediation $X \rightarrow K \rightarrow Y$. In particular, the first way is the straightforward causal effect. The reasons behind the second way causal effect are: 1) $X \rightarrow K$: the domain of a visual dataset is determined by the candidate image data, and thus the corresponding classification knowledge for the domain is determined by the candidate

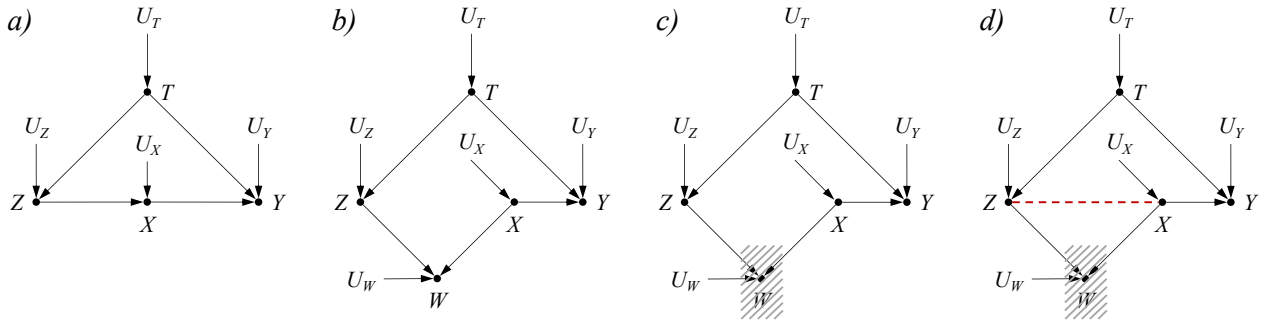

Figure 3: Examples of the graphical model of SCM. The subfigures a) and b) denote the common cases of performing the intervention on X to explore the causal effects between X and Y. The subfigure c) presents a specific case of, given W, performing the conditional intervention on X to explore the covariate-specific causal effects between X and Y. The red dashed line in the subfigure d) denotes the unordered dependency relationship between X and Z. Note that we differentiate the major variables used here and in Figure 2 to avoid confusion.

image data. 2) $K \rightarrow Y$: the target label can be predicted based on the specific classification knowledge.

$\boldsymbol{P \rightarrow K \rightarrow Y \leftarrow F \leftarrow P}$. We denote P as the open-set semantic knowledge of the large-scale pre-trained model and F as the general discriminant knowledge for distinguishing foreground and background objects. Y is jointly determined by P via the mediation ways, including $P \rightarrow K \rightarrow Y$ and $P \rightarrow F \rightarrow Y$. Specifically, 1) $P \rightarrow K$: the domain-specific knowledge is extracted from the open-set semantic knowledge. 2) $P \rightarrow F$: the knowledge learned by the pre-trained model contains the discriminant knowledge for distinguishing foreground and background objects, because, during pre-training, the input data of the model includes pairs of an image and the corresponding description (or label), and the description focuses on representing features of *foreground* objects so that the pre-trained model contains the general discriminant knowledge for distinguishing foreground and background objects. 3) $F \rightarrow Y$: in the target downstream task, i.e., FSOD, the mentioned general discriminant knowledge for distinguishing foreground and background objects is critical to determine the label, since "*background*" is a particular label in such an experimental setting. Concretely, for the FSOD task, the open-set semantic knowledge includes the general discriminant knowledge to distinguish foreground and background objects and the domain-specific classification knowledge to specifically classify the foreground objects. Therefore, the mediation causal effect $P \rightarrow K \rightarrow Y \leftarrow F \leftarrow P$ holds.

The intuition behind our assumption of Figure 2 b) is that as the domain is fixed for a specific downstream task, the target categories are constant so that the corresponding classification knowledge is determined, and our expected causal effect between X and Y needs to be quantified conditional on K. According to the conditional independence theorem of chains in SCM (Glymour, Pearl, and Jewell 2016), given K, the mediation causal path $X \rightarrow K \rightarrow Y$ is blocked, i.e., X and Y are *independent* conditional on K in $X \rightarrow K \rightarrow Y$. However, according to the conditional dependence theorem of colliders in SCM (Glymour, Pearl, and Jewell 2016), X and P are *dependent* conditional on K in $X \rightarrow K \leftarrow P$ so that if we directly measure the causal effect $X \rightarrow Y$, the quantified results may be biased due to P. We aim to apply the adjustment approach to quantify the causal effect $X \rightarrow Y$ based on the backdoor criterion, yet the common definition of the backdoor path does not apply to the specific SCM case in Figure 2 b).

3.3 Discussion on the Backdoor Path

Definition 3.1 ***(The Backdoor Criterion*** *(Glymour, Pearl, and Jewell 2016)) Given an ordered pair (X, Y) in a directed acyclic causal graph G, a set of variables Z satisfies the* ***backdoor criterion*** *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every* ***backdoor path*** *between X and Y having an arrow into X.*

According to Definition 3.1, (Glymour, Pearl, and Jewell 2016) proposes the common definition of the backdoor path to demarcate the scope of application of the backdoor criterion. Such a backdoor path definition can be applied in most cases, e.g., the common cases demonstrated in Figure 3 a) and b). However, the common backdoor path definition cannot be applied in the case of conditional intervention. For instance, as shown in Figure 3 c) and d), given W, we aim to explore the covariate-specific causal effects between X and Y. As the conditional dependence theorem of colliders in SCM, if a collider node, i.e., one node receiving edges from two other nodes, exists, conditioning on the collision node produces an unordered dependence between the node's parents. Therefore, the causal path $Y \leftarrow T \rightarrow Z \rightarrow W \leftarrow X$ originally blocked by the collider node, in Figure 3 b), is connected conditional on W, in Figure 3 c). According to the common backdoor path definition in Definition 3.1, the connected path $Y \leftarrow T \rightarrow Z \rightarrow W \leftarrow X$ is still not a backdoor path, but the confounder T impacts both X and Y so that the true covariate-specific causal effects between X and Y cannot be directly calculated.

To tackle this issue, recent works (Van der Zander, Liskiewicz, and Textor 2014; Perkovic et al. 2018; Correa and Bareinboim 2017) are committed to exploring *updated*

SCMs to determine how to impose the backdoor adjustment in different scenarios, yet they require building a new SCM by using more complex symbology and case-specific analyses. Inspired by such approaches, we propose to expand the backdoor criterion's application boundary on the conditional intervention cases without using extra symbols, which shares the intrinsic intuition with (Van der Zander, Liskiewicz, and Textor 2014; Perkovic et al. 2018; Correa and Bareinboim 2017). In detail, given an ordered pair of variables (X, Y) in a directed acyclic structural causal graph G, a path satisfies the definition of the backdoor path relative to (X, Y) if it contains a *confounder* T that jointly infers both X and Y, e.g., for T and X, T is the cause of X, or T and X are dependent if no direct causal relationship exists.

As shown in Figure 3 d), T has direct ordered causal relationships with Z and Y. Z and X are dependent without a direct causal relationship as denoted by the red dashed line in Figure 3 d). Therefore, T is the shared cause of X and Y. T can be treated as a *confounder*, and the path $Y \leftarrow T \rightarrow Z \rightarrow W \leftarrow X$ is a backdoor path conditional on W. We can achieve the true covariate-specific causal effects between X and Y by performing the backdoor adjustment.

3.4 Conditional Causal Intervention via Backdoor Adjustment

An ideal FSOD model should capture the true causality between X and Y and can generalize to unseen samples well. For the knowledge distillation empowered training approach, as shown in Figure 2, we expect to capture the direct causal relationship between X and Y independent of P. However, from the proposed SCM demonstrated in Figure 2 b), the increased likelihood of Y given X is not only due to $X \rightarrow Y$, but also the spurious correlation via $X \rightarrow K \leftarrow P \rightarrow F \rightarrow Y$ conditional on K. Consequently, the prediction of Y is based on not only the input data X, but also the semantic knowledge taught by the pre-trained model, which is demonstrated by the experiments in Figure 1. Therefore, to pursue the true causality between X and Y, we need to use the conditional causal intervention $P(Y(X)|do(X))$ instead of the $P(Y(X)|X)$.

We propose to use the backdoor adjustment (Glymour, Pearl, and Jewell 2016) to eliminate the interference of different teachers' knowledge. The backdoor adjustment assumes that we can observe and adjust the set of variables satisfying the backdoor criterion to achieve the true causal effect with intervention. In the proposed SCM, the semantic knowledge contained in P is immeasurable, because the input data domain is constant for a specific task. However, the general discriminant knowledge for distinguishing foreground and background objects contained in F is shared among different tasks so that we can observe and adjust F to achieve the true causality between X and Y. Formally, the backdoor adjustment for the proposed SCM is presented as:

$$P(Y(X)|do(X)) = \sum_{j=1}^{N^F} P(Y(X)|X, \hat{F}_j)P(\hat{F}_j), \qquad (2)$$

where $P(Y(X)|do(X))$ represents the true causality between X and Y, and $\hat{F}_j$ denotes the stratified knowledge of F, i.e., $F = \left\{\hat{F}_j | j \in [\![1, N^F]\!]\right\}$.

Algorithm 1: D&R Training and Fine-tuning Paradigm

Input:
#: N, minibatch size
#: f_{base}, f_{all}, detectors
#: λ, hyper-parameter, the weight of $\mathcal{L}_{\text{D\&R}}$
#: lr_{base}, lr_{all}, learning rates

$\#\ training\ on\ samples\ of\ base\ classes$
repeat
  Iteratively sample minibatch $X_{base} = \{X_i\}_{i=1}^{N}$.
  $\mathcal{L}_{base} \leftarrow \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RCNN}} + \mathcal{L}_{cross-entropy}^{Aux}$
  $f_{base} \leftarrow f_{base} - lr_{base} \nabla_f \mathcal{L}_{base}$
until f_{base} converge.
$\#\ fine\text{-}tuning\ on\ samples\ of\ all\ classes$
Initialize f_{all} with the weight of converged f_{base}.
repeat
  Iteratively sample minibatch $X_{all} = \{X_i\}_{i=1}^{N}$.
  $\mathcal{L}_{all} \leftarrow \mathcal{L}_{\text{RPN}} + \mathcal{L}_{\text{RCNN}} + \mathcal{L}_{cross-entropy}^{Aux} + \lambda \mathcal{L}_{\text{D\&R}}$
  $f_{all} \leftarrow f_{all} - lr_{all} \nabla_f \mathcal{L}_{all}$
until f_{all} converge.

4 Methodology

4.1 Overview

We provide the functional implementations in Figure 4. To illustrate our method more clearly, we present the training and fine-tuning paradigm of D&R in Algorithm 1. We adopt the two-stage training scheme following DeFRCN (Qiao et al. 2021). In the first stage, we train the detector with abundant samples of base classes. Besides the loss terms in DeFRCN, i.e., $\mathcal{L}_{\text{RPN}}$ and $\mathcal{L}_{\text{RCNN}}$, we introduce the conventional cross-entropy loss $\mathcal{L}_{cross-entropy}^{Aux}$ to guide the training of the feature extractor and the projector (refer to Figure 4). While fine-tuning the network with samples of all categories, i.e., base categories and novel categories, under the generalized few-shot object detection setting (G-FSOD), we propose $\mathcal{L}_{\text{D\&R}}$ to boost the performance of the main detection branch. We elaborate on details of $\mathcal{L}_{\text{D\&R}}$ in Sections 4.2 and 4.3. After training and fine-tuning, the teacher is abandoned, and the main branch is used to produce detection results.

4.2 Knowledge Distillation with Backdoor Adjustment

We present the implementation of the backdoor adjustment during the fine-tuning phase. As shown in Equation 2, we provide the detailed functional implementations for the knowledge distillation with backdoor adjustment as follows. The foundational idea behind the knowledge distillation is aligning the classification probabilities generated by the teacher and student models (the teacher model is fixed while the student model is trainable) in order to promote the student model to learn both "correct" and "incorrect" knowledge from the teacher model. Therefore, we represent the functional implementations of the $P(Y(X)|X, \hat{F}_j)$ by

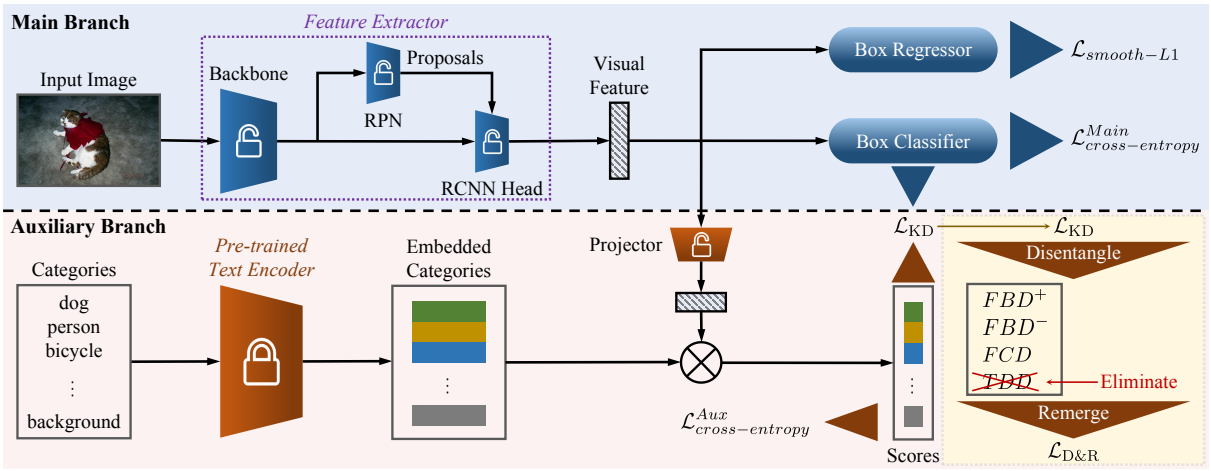

Figure 4: D&R's architecture. We introduce an auxiliary branch to the main branch of the FSOD benchmark approach. The d-dimension embedded categories are obtained from a frozen pre-trained text encoder. A linear projector is applied to transform the visual feature into d-dimension vectors. While training base classes, we backpropagate $\mathcal{L}^{Main}_{cross-entropy}$, $\mathcal{L}^{Aux}_{cross-entropy}$ and $\mathcal{L}_{smooth-L1}$ to train the feature extractor and projector in an end-to-end manner. We introduce $\mathcal{L}_{\text{D\&R}}$ to promote the detection during fine-tuning.

adopting the loss defined in Equation 1:

$$P(Y(X)|X,\hat{F}_j) = \sum_{i=1}^{N^C}\left(\lfloor p_i^{\mathcal{T}}\rceil^{\bar{K}} log\frac{\lfloor p_i^{\mathcal{T}}\rceil^{\bar{K}}}{\lfloor p_i^{\mathcal{S}}\rceil^{\bar{K}}} + \lfloor p_i^{\mathcal{T}}\rceil^{\hat{F}_j} log\frac{\lfloor p_i^{\mathcal{T}}\rceil^{\hat{F}_j}}{\lfloor p_i^{\mathcal{S}}\rceil^{\hat{F}_j}}\right), \quad (3)$$

where $\bar{K}$ denotes the restricted classification knowledge extracted from K due to the data domain of a specific task. $\lfloor p_i^{\mathcal{T}}\rceil^{\bar{K}}$ and $\lfloor p_i^{\mathcal{S}}\rceil^{\bar{K}}$ denote the distillation of the knowledge related to $\bar{K}$, and $\lfloor p_i^{\mathcal{T}}\rceil^{\hat{F}_j}$ and $\lfloor p_i^{\mathcal{S}}\rceil^{\hat{F}_j}$ denote the distillation of the stratified knowledge related to F in SCM. As a result, we implement the overall backdoor adjustment by

$$P(Y(X)|do(X)) = \sum_{i=1}^{N^C}\sum_{j=1}^{N^F}\left(\lfloor p_i^{\mathcal{T}}\rceil^{\bar{K}} log\frac{\lfloor p_i^{\mathcal{T}}\rceil^{\bar{K}}}{\lfloor p_i^{\mathcal{S}}\rceil^{\bar{K}}} + \lfloor p_i^{\mathcal{T}}\rceil^{\hat{F}_j} log\frac{\lfloor p_i^{\mathcal{T}}\rceil^{\hat{F}_j}}{\lfloor p_i^{\mathcal{S}}\rceil^{\hat{F}_j}}\right). \quad (4)$$

According to the proposed SCM, the restricted classification knowledge $\bar{K}$ is related to the specific FSOD task so that $\bar{K}$ is naturally contained in the adjusted knowledge distillation objective, i.e., Equation 4. Following the principle of backdoor adjustment, we further sum up all conditional causal effects based on adjustment for the stratified knowledge $\hat{F}_j$ in Equation 4. Therefore, we *disentangle* the knowledge distillation objective and eliminate the *confounder* term and then *remerge* the remaining terms according to the proposed backdoor adjustment methodology.

4.3 Disentangle and Remerge

Inspired by (Zhao et al. 2022), we disentangle the knowledge distillation objective for FSOD into four terms: 1) two terms for positive samples, i.e., the labels of target samples belong to the foreground categories, including *positive Foreground and Background knowledge Distillation* (FBD$^+$) and *Target Discrimination knowledge Distillation* (TDD); 2) one term for negative samples, where the shared label of target samples is "background", i.e., *negative Foreground and Background knowledge Distillation* (FBD$^-$); 3) a common term for all samples, i.e., *Foreground Classification knowledge Distillation* (FCD).

FBD$^+$ and FBD$^-$ present the distillation objective of stratified knowledge $\hat{F}$ extracted from general discriminant knowledge for distinguishing foreground and background objects F, respectively. Specifically, FBD$^+$ is the disentangled knowledge distillation objective to measure the similarity between the teacher's and student's binary probabilities of the "background" category and *non-target* foreground category group. FBD$^-$ is the objective to measure the similarity between the teacher's and student's binary probabilities of the *target* "background" category and foreground category group. TDD is the considered confounder, which denotes the objective to measure the similarity between the teacher's and student's binary probabilities of the *target* category and the *non-target* category group. Our explanation of the observation of Figure 1 states that in the process of knowledge distillation, the empirical error of the teacher model degenerates the student model's prediction of the target labels, since there exists a confounder leading the student model to learn the exceptional correlation relationship between the classification knowledge and general discriminant knowledge for distinguishing foreground and background objects of the teacher model during knowledge distillation (see Section 1 for details). FCD represents the distillation objective of the restricted classification knowledge $\bar{K}$, which is the objective to measure the similarity between the teacher's and student's multiple probabilities among *non-*

Methods	Novel Set 1					Novel Set 2					Novel Set 3				
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Results of single run, following TFA (Wang et al. 2020)															
FSRW (Kang et al. 2019)	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
TFA (Wang et al. 2020)	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR (Wu et al. 2020)	41.7	42.5	51.4	55.2	61.8	24.4	29.3	39.2	39.9	47.8	35.6	41.8	42.3	48.0	49.7
FSCE (Sun et al. 2021)	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
SRR-FSD‡ (Zhu et al. 2021)	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
Meta Faster R-CNN (Han et al. 2022a)	43.0	54.5	60.6	**66.1**	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6
FCT (Han et al. 2022b)	49.9	57.1	57.9	63.2	**67.1**	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7
Kaul, Xie, and Zisserman (2022)	54.5	53.2	58.8	63.2	65.7	32.8	29.2	**50.7**	49.8	50.6	48.4	52.7	55.0	59.6	59.6
DeFRCN* (Qiao et al. 2021)	55.1	61.9	64.9	65.8	66.2	33.8	45.1	46.1	**53.2**	52.3	51.0	56.6	55.6	59.7	**61.9**
D&R (Ours)‡	**60.4**	**64.0**	**65.2**	64.7	66.3	**37.9**	**46.8**	48.1	52.7	**53.1**	**55.7**	**57.9**	**57.6**	**60.6**	**61.9**
Average results of 30 runs, following TFA (Wang et al. 2020)															
FRCN+ft-full (Yan et al. 2019)	9.9	15.6	21.6	28.0	35.6	9.4	13.8	17.4	21.9	29.8	8.1	13.9	19.0	23.9	31.0
Xiao et al (Xiao and Marlet 2020)	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6
TFA (Wang et al. 2020)	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
FSCE (Sun et al. 2021)	32.9	44.0	46.8	52.9	59.7	23.7	30.6	38.4	43.0	48.5	22.6	33.4	39.5	47.3	54.0
DCNet (Hu et al. 2021)	33.9	37.4	43.7	51.1	59.6	23.2	24.8	30.6	36.7	46.6	32.3	34.9	39.7	42.6	50.7
FCT (Han et al. 2022b)	38.5	49.6	53.5	59.8	64.3	25.9	34.2	40.1	44.9	47.4	34.7	43.9	49.3	53.1	56.3
DeFRCN* (Qiao et al. 2021)	39.3	50.9	55.3	**61.8**	65.3	27.4	36.8	40.4	45.1	50.8	35.0	45.1	50.2	55.7	58.9
D&R (Ours)‡	**41.0**	**51.7**	**55.7**	**61.8**	**65.4**	**30.7**	**39.0**	**42.5**	**46.6**	**51.7**	**37.9**	**47.1**	**51.7**	**56.8**	**59.5**

Table 1: FSOD results (%) on VOC. * denotes the method re-implemented with one single GPU. ‡indicates the methods using external knowledge.

target foreground categories.

To perform the expected knowledge distillation with backdoor adjustment, we eliminate the confounder TDD and further remerge the remaining disentangled objective terms. According to Equation 1 and Equation 4, we derive the final loss function for the proposed D&R:

$$\begin{aligned}\mathcal{L}_{\text{D\&R}} =& \alpha \text{KL}\left(\mathcal{P}^{\mathcal{T}}_{\text{FBD}^+} \| \mathcal{P}^{\mathcal{S}}_{\text{FBD}^+}\right) \\ &+ \beta \text{KL}\left(\mathcal{P}^{\mathcal{T}}_{\text{FBD}^-} \| \mathcal{P}^{\mathcal{S}}_{\text{FBD}^-}\right) + \text{KL}\left(\mathcal{P}^{\mathcal{T}}_{\text{FCD}} \| \mathcal{P}^{\mathcal{S}}_{\text{FCD}}\right) \\ =& \sum_{i=1}^{N^C} \left(\alpha \lfloor p_i^{\mathcal{T}} \rceil^{\hat{F}_1} log \frac{\lfloor p_i^{\mathcal{T}} \rceil^{\hat{F}_1}}{\lfloor p_i^{\mathcal{S}} \rceil^{\hat{F}_1}} + \beta \lfloor p_i^{\mathcal{T}} \rceil^{\hat{F}_2} log \frac{\lfloor p_i^{\mathcal{T}} \rceil^{\hat{F}_2}}{\lfloor p_i^{\mathcal{S}} \rceil^{\hat{F}_2}} \right. \\ &\left. + \lfloor p_i^{\mathcal{T}} \rceil^{\bar{K}} log \frac{\lfloor p_i^{\mathcal{T}} \rceil^{\bar{K}}}{\lfloor p_i^{\mathcal{S}} \rceil^{\bar{K}}} \right),\end{aligned} \tag{5}$$

where $\hat{F}_1$ and $\hat{F}_2$ denote the stratified knowledge of F, corresponding to FBD^+ and FBD^-, respectively. α and β are coefficients that control the impact of the terms for positive samples and negative samples in knowledge distillation.

5 Experiments

5.1 Experimental Setting

Benchmarks. We benchmark D&R on Pascal VOC (Everingham et al. 2010) and COCO (Lin et al. 2014) datasets following the previous experimental settings (Wang et al. 2020; Qiao et al. 2021) for a fair comparison. For Pascal VOC, 15 classes are randomly selected as base classes, and the remaining 5 classes are novel classes. Each novel class has $K = 1, 2, 3, 5, 10$ annotated training samples. We train the network with VOC07 and VOC12 train/val set, and evaluate our method with VOC07 test set using AP_{50} as the evaluation metric. For COCO, there are 60 base categories that are disjoint with VOC and 20 novel classes. Each novel class has $K = 1, 2, 3, 5, 10, 30$ samples. We report COCO-style mAP of novel classes for COCO.

Implementation Details. Our model is built upon the state-of-the-art method DeFRCN (Qiao et al. 2021) with a backbone network ResNet-101. We use SGD as the optimizer with a batch size of 8. All models are trained with a single GPU. Due to the change in batch size, the number of training iterations is doubled based on the implementation of DeFRCN, and the learning rate is halved. Other parameters are exactly the same as DeFRCN. We add four additional hyper-parameters. For the experiments of COCO, the distillation temperature is 5, and the weight of the distillation loss is 5. α and β in $\mathcal{L}_{\text{D\&R}}$ are 4 and 0.5, respectively. For the experiments on Pascal VOC, we set the temperature to 10 and the distillation loss weight to 1 empirically. α and β are 10 and 2. Moreover, the loss terms of positive and negative samples are averaged separately in the calculation of $\mathcal{L}_{\text{D\&R}}$ to achieve a balance. Our method is implemented on Pytorch 1.9. Results on the COCO dataset are obtained using a single Tesla V100 GPU with a memory of 32G. The experiments on the VOC dataset are conducted using one Geforce RTX3090 GPU with a memory of 24G. The operating system we use is Ubuntu18.04.

5.2 Comparison Results

We report AP_{50} for novel classes on three data splits of Pascal VOC in Table 1, and the results of COCO-style mAP

Methods	Shot Number					
	1	2	3	5	10	30
Results of single run, following TFA (Wang et al. 2020)						
FSRW (Kang et al. 2019)	-	-	-	-	5.6	9.1
TFA (Wang et al. 2020)	3.4	4.6	6.6	8.3	10.0	13.7
MPSR (Wu et al. 2020)	2.3	3.5	5.2	6.7	9.8	14.1
FSCE (Sun et al. 2021)	-	-	-	-	11.9	16.4
SRR-FSD‡ (Zhu et al. 2021)	-	-	-	-	11.3	14.7
FCT (Han et al. 2022b)	5.6	7.9	11.1	14.0	17.1	21.4
DeFRCN (Qiao et al. 2021)	6.5	11.8	13.4	15.3	18.6	**22.5**
D&R (Ours)‡	**8.3**	**12.7**	**14.3**	**16.4**	**18.7**	21.8
Average results of 10 runs, following TFA (Wang et al. 2020)						
FRCN+ft-full	1.7	3.1	3.7	4.6	5.5	7.4
TFA	1.9	3.9	5.1	7.0	9.1	12.1
DeFRCN	4.8	8.5	10.7	13.5	**16.7**	**21.0**
D&R (Ours)‡	**6.1**	**9.5**	**11.5**	**13.9**	16.4	20.0

Table 2: FSOD results (%) on COCO. ‡indicates the methods using external knowledge.

are shown in Table 2 for COCO. We observe that D&R achieves state-of-the-art performance on most tasks. Especially, D&R has more impressive improvements with fewer annotated samples. At higher shots, D&R can also obtain competitive results. Specifically, D&R is, on average, 1.58% higher than the best baseline method on the VOC dataset. D&R averagely beats the best benchmark method by 0.68% on the COCO dataset. COCO is a larger dataset, and Pascal VOC is smaller with respect to the category number and the sample size. Benchmark approaches and D&R achieve relatively consistent performance on COCO, and as shown in Table 2, D&R's improvements are more consistent.

Furthermore, we report the average results of multiple repeated runs over different training samples. For Pascal VOC, as the few-shot detection performance on VOC is quite unstable, we set a fixed random seed for all experiments to stabilize the results. Moreover, results reported in DeFRCN (Qiao et al. 2021) are produced with multi-GPUs. As the results are extremely different when a single GPU is used, we re-produce the results of DeFRCN (Qiao et al. 2021) with one GPU based on the officially released base model and mark the results in Table 1 with *. From Table 1, we learn that our proposed D&R improves the performance by 2.6%, 1.7%, 1.3%, 0.9%, and 0.5% on average when the shot number $K = 1, 2, 3, 5, 10$, respectively. For COCO, we report the average results of 10 repeated runs of different training samples in Table 2. At lower shots, our D&R has consistent improvements compared with the baseline method DeFRCN (Qiao et al. 2021). When the shot number is 10 or higher, knowledge from CLIP cannot provide much help. This is in agreement with the results on VOC.

5.3 Ablation Study

Effectiveness of $\mathcal{L}_{\text{D\&R}}$. We conduct the ablation experiments to take a closer look at D&R in Table 3. We find that D&R, shown in the last row, outperforms all ablation variants, and the performance gains mainly owe to the

Components					Shot Number					
KD	TDD	FBD^+	FBD^-	FCD	1	2	3	5	10	30
					6.77	10.94	12.96	15.22	18.02	21.63
✓					7.69	11.96	13.73	15.82	18.59	**21.84**
		✓	✓		7.71	11.89	13.62	15.53	18.11	21.41
				✓	8.10	12.33	14.04	16.20	**18.70**	21.76
	✓	✓	✓	✓	8.05	12.40	14.26	16.34	**18.70**	21.67
			✓	✓	7.91	12.22	13.94	16.07	18.59	21.72
		✓		✓	8.25	12.67	14.10	**16.50**	18.55	21.81
		✓	✓	✓	**8.29**	**12.71**	**14.27**	16.43	18.65	21.82

Table 3: FSOD results (%) of the ablation experiments on COCO. The first row indicates the model without the knowledge distillation, and the second row indicates the model with the vanilla knowledge distillation.

reliable knowledge provided by $\mathcal{L}_{\text{D\&R}}$, which proves the effectiveness of D&R. Although the FBD^- + FCD variant even underperforms the sole FCD variant, according to the backdoor adjustment-based approach, FBD^- can improve the performance of our model (as shown in the last row). Such observations demonstrate the effectiveness of the proposed backdoor adjustment-based learning paradigm. In most cases, combining FBD^+, FBD^-, and FCD achieves preferable results, but adding TDD may degenerate the performances, e.g., comparing the fifth row and the last row, we observe that TDD indeed degenerates the model's performance, which proves our statement treating TDD as an unexpected confounder, thereby demonstrating the empirical effectiveness of D&R. Regarding the results, we have a further observation: as the shot number grows, the improvement brought by the distillation loss becomes limited, including $\mathcal{L}_{\text{D\&R}}$. The reason is that the semantic information of novel classes is extremely scarce at lower shots, so the knowledge from pre-trained models can effectively improve models. However, when more training samples are available, the knowledge distillation provides less additional information, and the performance improvement is limited.

Analysis in Training Consumption. On COCO, during base training, DeFRCN needs 0.834s/iter with a memory of 10159M, while D&R costs 0.839s/iter with 10179M. For fine-tuning, DeFRCN costs 0.749s/iter with 10008M, and D&R costs 0.794s/iter with 10040M. The inference time of both methods is 0.08s/image. The training iterations are the same. The results on VOC are consistent. Overall, our approach adds negligible time and space consumption during training and brings no extra consumption during testing.

Hyper-Parameters There are four important hyper-parameters in $\mathcal{L}_{\text{D\&R}}$: distillation temperature T, distillation loss weight λ, and two parameters α and β for the weights of FBD^+ and FBD^- in $\mathcal{L}_{\text{D\&R}}$, respectively. We analyze the effectiveness of different hyper-parameters one by one.

Distillation Temperature. We distill the FSOD model using different temperatures with vanilla knowledge distillation on the COCO dataset. As shown in Table 4, temperature 5 achieves the best or the second-best performance in

Temperature	Shot Number					
	1	2	3	5	10	30
1	7.00	11.10	13.18	15.32	17.85	21.59
5	**7.31**	**11.56**	**13.49**	15.65	18.13	21.66
10	7.16	11.27	13.37	**15.74**	18.09	**21.93**
20	6.95	11.43	13.22	15.45	**18.27**	21.66

Table 4: Comparison results (%) to evaluate the effectiveness of different distillation temperatures on the COCO dataset.

Weight of KD	Shot Number					
	1	2	3	5	10	30
1	7.31	11.56	13.49	15.65	18.13	21.66
5	**7.69**	**11.96**	**13.73**	**15.82**	**18.59**	**21.84**
10	7.56	11.95	13.54	15.71	18.24	21.65
20	7.11	11.63	13.13	14.89	18.08	21.16

Table 5: FSOD results (%) of the effectiveness of different distillation loss weights on the COCO dataset.

all shots, so we set the distillation temperature to 5 in all following experiments for COCO.

Distillation Loss Weight. With temperature 5, we test the impact of different distillation loss weights. As shown in Table 5, the model with weight 5 performs best in all cases, so we select 5 as the weight of distillation loss empirically. Meanwhile, nice results are achieved with weight 10, so we argue that the performance is relatively robust to different weights of distillation loss.

Weights for FBD. As shown in Equation 5, there are two parameters to be tuned, α and β. We carefully explore the impact of their different values and illustrate the results in Figure 5. We set α to 4 and β to 0.5, which achieves the best result.

5.4 Discussion on Knowledge Distillation Variants

As shown in Figure 4, D&R adopts the large-scale pre-trained text encoder as the teacher. The reasons include 1) "text" is the semantically-dense data while "image" is the semantically-sparse data so that the text encoder teacher can more efficiently improve the student model to explore semantic information from the input data (including images and category texts) than the image encoder teacher, and such efficiency is crucial to the FSOD, which requires to train the model with a few data; 2) the issue of domain shift is far more serious for image data than for text data. Thus, even with the limited data, text encoders can still achieve consistent performance; 3) for vision-language models, the captured knowledge is shared between the text and image encoders, which is due to the training paradigm (Radford et al. 2021). Therefore, the text encoder of vision-language models can teach the student model the general knowledge to capture the semantic knowledge from images; 4) the time and space complexities of adopting the image encoder as the teacher is excessively larger than adopting the text model as the teacher, which demonstrates that it is not worthy of

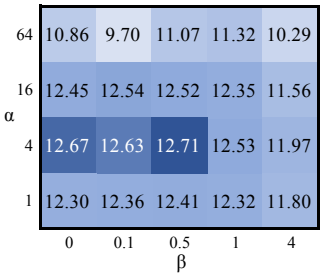

Figure 5: Experimental results (%) of the hyper-parameter comparisons of α and β.

Teacher Model	Complexity		Shot Number				
	Time	Space	1	2	3	5	10
Word2Vec	0.7s/iter	9.6G	56.8	63.0	64.3	64.1	66.2
GloVe	0.7s/iter	9.6G	56.7	63.5	64.0	63.7	65.8
CLIP (I+T)	2.0s/iter	14.2G	58.7	**64.1**	**65.6**	**65.9**	66.5
CLIP (T)	0.7s/iter	9.6G	**59.1**	63.4	65.3	65.0	**66.8**

Table 6: FSOD results (%) of the comparisons of our proposed method adopting different teachers on VOC Novel Set 1. The *underlines* denote the second best results. The complexities are measured during training.

adopting the image encoder as the teacher.

To prove our statements above, we conduct explorations of D&R by using different teacher variants for knowledge distillation, including CLIP(I+T) having both CLIP image and text encoders as the teachers, CLIP(T) only having the CLIP text encoder, Word2Vec, and GloVe. Note that the last three variants only have the pre-trained text encoder as the teacher. From Table 6, we observe that as the comparison between CLIP(T) and CLIP(I+T), the improvement provided by the CLIP image encoder is limited, but the additional time and space consumption is not negligible. Additionally, the main model of D&R, i.e., CLIP(T), achieves top-2 performance on most tasks. The empirical results demonstrate the proposed statement and D&R's effectiveness and efficiency.

6 Conclusion

We introduce the knowledge distillation to FSOD tasks. Then, we discover that the empirical error of the teacher model degenerates the prediction performance of the student model. To tackle this latent flaw, we develop a Structural Causal Model and propose a backdoor adjustment-based knowledge distillation method, D&R. Empirically, D&R outperforms state-of-the-art methods on multiple benchmark datasets.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments. This work is supported by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDA19020500. The authors are grateful to Hong Wu for the fruitful inspiration.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. In *ICCV*, 2425–2433.

Bucila, C.; Caruana, R.; and Niculescu-Mizil, A. 2006. Model compression. In *Proc. of KDD*, 535–541.

Chang, C.; Adam, G.; and Goldenberg, A. 2021. Towards Robust Classification Model by Counterfactual and Invariant Data Generation. In *Proc. of CVPR*, 15212–15221.

Chen, W.; Liu, Y.; Kira, Z.; Wang, Y. F.; and Huang, J. 2019a. A Closer Look at Few-shot Classification. In *Proc. of ICLR*.

Chen, Y.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019b. UNITER: Learning UNiversal Image-TExt Representations. *CoRR*.

Cho, J. H.; and Hariharan, B. 2019. On the Efficacy of Knowledge Distillation. In *Proc. of ICCV*, 4793–4801.

Correa, J.; and Bareinboim, E. 2017. Causal effect identification by adjustment under confounding and selection biases. In *Proc. of AAAI*.

Everingham, M.; Gool, L. V.; Williams, C. K. I.; Winn, J. M.; and Zisserman, A. 2010. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*, 303–338.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proc. of ICML*, 1126–1135.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Han, G.; Huang, S.; Ma, J.; He, Y.; and Chang, S. 2022a. Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment. In *Proc. of AAAI*, 780–789.

Han, G.; Ma, J.; Huang, S.; Chen, L.; and Chang, S.-F. 2022b. Few-Shot Object Detection With Fully Cross-Transformer. In *Proc. of CVPR*, 5321–5330.

He, Y.; Shen, Z.; and Cui, P. 2021. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, 107383.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*.

Hu, H.; Bai, S.; Li, A.; Cui, J.; and Wang, L. 2021. Dense Relation Distillation With Context-Aware Aggregation for Few-Shot Object Detection. In *Proc. of CVPR*, 10185–10194.

Huang, L.; Wang, W.; Chen, J.; and Wei, X. 2019. Attention on Attention for Image Captioning. In *ICCV*, 4633–4642.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proc. of ICML*.

Kang, B.; Liu, Z.; Wang, X.; Yu, F.; Feng, J.; and Darrell, T. 2019. Few-Shot Object Detection via Feature Reweighting. In *Proc. of ICCV*, 8419–8428.

Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R. S.; Giryes, R.; and Bronstein, A. M. 2019. RepMet: Representative-Based Metric Learning for Classification and Few-Shot Object Detection. In *Proc. of CVPR*, 5197–5206.

Kaul, P.; Xie, W.; and Zisserman, A. 2022. Label, Verify, Correct: A Simple Few Shot Object Detection Method. In *Proc. of CVPR*, 14237–14247.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; and Gao, J. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *Proc. of ECCV*, 121–137.

Lin, T.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of ECCV*, 740–755.

Lopez-Paz, D.; Nishihara, R.; Chintala, S.; Schölkopf, B.; and Bottou, L. 2017. Discovering Causal Signals in Images. In *Proc. of CVPR*, 58–66.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 13–23.

Lv, F.; Liang, J.; Li, S.; Zang, B.; Liu, C. H.; Wang, Z.; and Liu, D. 2022. Causality Inspired Representation Learning for Domain Generalization. In *Proc. of CVPR*, 8036–8046.

Ma, Z.; Luo, G.; Gao, J.; Li, L.; Chen, Y.; Wang, S.; Zhang, C.; and Hu, W. 2022. Open-Vocabulary One-Stage Detection with Hierarchical Visual-Language Knowledge Distillation. In *Proc. of CVPR*, 14074–14083.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NeurIPS*, 3111–3119.

Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 96–146.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *Proc. of EMNLP*, 1532–1543.

Perkovic, E.; Textor, J.; Kalisch, M.; and Maathuis, M. H. 2018. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs.

Qiao, L.; Zhao, Y.; Li, Z.; Qiu, X.; Wu, J.; and Zhang, C. 2021. DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection. In *Proc. of ICCV*, 8661–8670.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. of ICML*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2015. FitNets: Hints for Thin Deep Nets. In *Proc. of ICLR*.

Snell, J.; Swersky, K.; and Zemel, R. S. 2017. Prototypical Networks for Few-shot Learning. In *Proc. of NeurIPS*, 4077–4087.

Sun, B.; Li, B.; Cai, S.; Yuan, Y.; and Zhang, C. 2021. FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding. In *Proc. of CVPR*, 7352–7362.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H. S.; and Hospedales, T. M. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *Proc. of CVPR*, 1199–1208.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP/IJCNLP (1)*, 5099–5110.

Tung, F.; and Mori, G. 2019. Similarity-Preserving Knowledge Distillation. In *Proc. of ICCV*, 1365–1374.

Van der Zander, B.; Liskiewicz, M.; and Textor, J. 2014. Constructing Separators and Adjustment Sets in Ancestral Graphs. In *CI@ UAI*, 11–24.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching Networks for One Shot Learning. In *Proc. of NeurIPS*, 3630–3638.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal Attention for Unbiased Visual Recognition. In *Proc. of ICCV*, 3071–3080.

Wang, X.; Huang, T. E.; Gonzalez, J.; Darrell, T.; and Yu, F. 2020. Frustratingly Simple Few-Shot Object Detection. In *Proc. of ICML*, 9919–9928.

Wu, J.; Liu, S.; Huang, D.; and Wang, Y. 2020. Multi-scale Positive Sample Refinement for Few-Shot Object Detection. In *Proc. of ECCV*, 456–472.

Xiao, Y.; and Marlet, R. 2020. Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild. In *Proc. of ECCV*, 192–210.

Yan, X.; Chen, Z.; Xu, A.; Wang, X.; Liang, X.; and Lin, L. 2019. Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning. In *Proc. of ICCV*, 9576–9585.

Yang, C.; Xie, L.; Qiao, S.; and Yuille, A. L. 2019. Training deep neural networks in generations: A more tolerant teacher educates better students. In *Proc. of AAAI*, 5628–5635.

Yim, J.; Joo, D.; Bae, J.; and Kim, J. 2017. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *Proc. of CVPR*, 7130–7138.

You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning with Semantic Attention. In *CVPR*, 4651–4659.

Yue, Z.; Wang, T.; Sun, Q.; Hua, X.; and Zhang, H. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. In *Proc. of CVPR*, 15404–15414.

Zagoruyko, S.; and Komodakis, N. 2017. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. In *Proc. of ICLR*.

Zhang, W.; and Wang, Y. 2021. Hallucination Improves Few-Shot Object Detection. In *Proc. of CVPR*, 13008–13017.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *CoRR*.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled Knowledge Distillation. In *Proc. of CVPR*, 11953–11962.

Zhao, X.; Zou, X.; and Wu, Y. 2021. Morphable Detector for Object Detection on Demand. In *Proc. of ICCV*, 4751–4760.

Zhu, C.; Chen, F.; Ahmed, U.; Shen, Z.; and Savvides, M. 2021. Semantic Relation Reasoning for Shot-Stable Few-Shot Object Detection. In *Proc. of CVPR*, 8782–8791.