



# Article Ensemble Transductive Propagation Network for Semi-Supervised Few-Shot Learning

Xueling Pan<sup>1,2</sup>, Guohe Li<sup>1,2,\*</sup> and Yifeng Zheng<sup>3,4</sup>

- Beijing Key Lab of Petroleum Data Mining, Department of Geophysics, China University of Petroleum, Beijing 102249, China; 2020310415@student.cup.edu.cn
- <sup>2</sup> College of Information Science and Engineering, China University of Petroleum, Beijing 102249, China
- <sup>3</sup> School of Computer Science, Minnan Normal University, Zhangzhou 363000, China; zhengyifengja@gmail.com
- <sup>4</sup> Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou 363000, China
- \* Correspondence: lgh202310@163.com

Abstract: Few-shot learning aims to solve the difficulty in obtaining training samples, leading to high variance, high bias, and over-fitting. Recently, graph-based transductive few-shot learning approaches supplement the deficiency of label information via unlabeled data to make a joint prediction, which has become a new research hotspot. Therefore, in this paper, we propose a novel ensemble semisupervised few-shot learning strategy via transductive network and Dempster-Shafer (D-S) evidence fusion, named ensemble transductive propagation networks (ETPN). First, we present homogeneity and heterogeneity ensemble transductive propagation networks to better use the unlabeled data, which introduce a preset weight coefficient and provide the process of iterative inferences during transductive propagation learning. Then, we combine the information entropy to improve the D-S evidence fusion method, which improves the stability of multi-model results fusion from the preprocessing of the evidence source. Third, we combine the L2 norm to improve an ensemble pruning approach to select individual learners with higher accuracy to participate in the integration of the few-shot model results. Moreover, interference sets are introduced to semi-supervised training to improve the anti-disturbance ability of the mode. Eventually, experiments indicate that the proposed approaches outperform the state-of-the-art few-shot model. The best accuracy of ETPN increases by 0.3% and 0.28% in the 5-way 5-shot, and by 3.43% and 7.6% in the 5-way 1-shot on miniImagNet and tieredImageNet, respectively.

**Keywords:** few-shot learning; meta learning; graph semi-supervision; label propagation; Gaussian kernel function; D-S evidence theory

# 1. Introduction

Deep learning is widely used in many practical applications, such as speech recognition [1], computer vision [2], and semantic segmentation [3]. It is data-driven and relies on a large amount of labeled data to train a model. However, in some scenarios, labeled data are costly to obtain. Therefore, how to utilize the limited labeled data to construct a reliable model is very important. Nowadays, inspired by human learning and utilizing prior knowledge to learn new concepts via only a handful of examples, few-shot learning (FSL) has attracted much more attention.

Few-shot learning has been divided into three categories, including augmentation, metric learning, and meta-learning. FSL usually adopts an episodic training mode. Each episodic training includes a support and a query set. The support set is constructed by randomly selecting *K* categories and *N* samples in each selected category from training data, namely the *K*-way *N*-shot. The query set is also randomly sampled from the *K* categories, but it has no intersection with the support set.



**Citation:** Pan, X.; Li, G.; Zheng, Y. Ensemble Transductive Propagation Network for Semi-Supervised Few-Shot Learning. *Entropy* **2024**, *26*, 135. https://doi.org/10.3390/ e26020135

Academic Editors: Shuangming Yang, Badong Chen, Shujian Yu and Luis Gonzalo Sánchez Giraldo

Received: 31 December 2023 Revised: 27 January 2024 Accepted: 30 January 2024 Published: 31 January 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). For augmentation approaches, they aim to increase the number of training samples or features to enhance data diversity. However, some basic augmentation operations need to be improved in the process of model training, such as rotating, flipping, cropping, translating, and adding noise into images [4,5]. With the development of deep learning, more sophisticated algorithms customized for FSL were proposed. Dual TriNet mapped the multi-level image feature into a semantic space to enhance the semantic vectors using the semantic Gaussian or neighborhood approaches [6]. ABS-Net established a repository of attribute features by conducting attribute learning on the auxiliary dataset to synthesize pseudo feature representations automatically [7].

Metric learning is designed to learn a pairwise similarity metric by exploiting the similarity information between samples. It means a similar sample pair has a high similarity score and vice versa. Matching Nets performed full context embeddings by adding external memories to extract features. It measures the similarity between samples via the cosine distance [8]. Proto Net constructed a metric space by computing distances between the prototype representations and test examples [9]. AM3 incorporated extra semantic representations into Proto Net [10]. TSCE utilized the mutual information maximization and ranking-based embedding alignment mechanism to implement knowledge transfer across domains, which maintains the consistency between the semantic and shared spaces, respectively [11]. Moreover, TSVR made the source and target domains have the same label space to quantify domain discrepancy by predicting the similarity/dissimilarity labels for semantic-visual fusions [12]. K-tuplet Nets changed the NCA loss of Proto Net into a K-tuplet metric loss [13]. The drawback of the above algorithms is that they cannot learn enough transferable knowledge in a small number of samples to enhance the model's performance.

Meta-learning approaches aim to utilize the transferring experience of the meta-learner to optimize a base learner. It is divided into three categories: Learn-to-Measure [14–16], Learn-to-Finetune [17,18], and Learn-to-Parameterize [19–21]. MAML learned a suitable initialization parameter via a multi-task training strategy to guarantee its generalization [22]. TAML utilized an entropy-maximization reduction to address the over-fitting problem [23]. DAE employed a graph neural network based on a denoising auto-encoder to generate model parameters [19]. However, the above solutions should further consider the related information between the support set and the query set. Even more importantly, learning a base learner for few-shot tasks is easy to overfit, which results in high-variance or lowconfidence predictions by lacking training data [24].

Nowadays, some researchers focus on measuring the relations between the support and the query instances via transductive graph theory. The transductive graph-based approaches [25–27] can effectively obtain the labels of the query set based on a few labeled samples. The main idea is that, regarding the samples of the support set and the query set as graph nodes, the nearest neighbor relationship between the support set and the query set is utilized for joint prediction to supplement the lack of label information. TPN employed the Gaussian kernel function to calculate the similarity as the weight to build a k-nearest neighbor graph (KNN-Graph), which uses the label propagation algorithm to transductively propagate labels between the support and query examples [25]. The drawback is that it may divide all graph vertices into a vast community or trap them in a local maximum to affect the stability and robustness of a model.

To address the above problems, we propose an Ensemble Transductive Propagation Network (ETPN). Firstly, two types of ensemble strategies are proposed, based on homogeneous and heterogeneous algorithms. These are referred to as Ho-ETPN (Homogeneous Ensemble Transductive Propagation Network) and He-ETPN (Heterogeneous Ensemble Transductive Propagation Network), respectively. Transductive inference, based on a graph, is used to extract valuable information shared between support-query pairs for label prediction. This approach circumvents the intermediate problem of defining a prediction function on an entire space in inductive learning. Secondly, a novel fusion strategy is proposed, based on an improved D-S evidence theory, to enhance the robustness of our proposal. The improved D-S evidence fusion approach first uses the Bhattacharyya distance to construct a conflict matrix between the mass function, and then uses this conflict matrix to obtain the support matrix. It then combines information entropy to recalculate the mass weight, realizing the pre-processing of the evidence source. This enhances the robustness and stability of few-shot classification. Thirdly, we propose an improved ensemble pruning approach to select individual learners with higher accuracy to participate in the integration of the few-shot model results. It employs the L2 norm to make the model more stable to small changes in the input and improve the model's robustness.

In summary, the key contributions of our approaches are summarized as follows:

- 1. *Ensemble framework*: Based on the individual graph learner framework, we propose two ensemble strategies including the homogeneous and heterogeneous models, named Ho-ETPN and He-ETPN, respectively. Moreover, during transductive propagation learning, we add the preset weight coefficient and give the process of iterative inferences.
- 2. *Ensemble pruning*: Proposing an improved ensemble pruning method to conduct the selective results fusion by screening the individual learner with higher accuracy.
- 3. *Combination strategy*: An improved D-S evidence aggregation method is proposed for comprehensive evaluation. To the best of our knowledge, it is the first work that explicitly considers the D-S evidence theory in few-shot learning.
- 4. *Effectiveness*: Extension experiences about supervised and semi-supervised conducted on miniImageNet and tieredImageNet datasets show that our solution yields competitive results on a few-shot classification. More challenging is that distracted classes are introduced during the process of the semi-supervised experiment.

#### 2. Related Work

(1) Transductive Graph Few-shot Learning

Recently, few-shot learning has become one of the hot spots. Transductive inference employs the valuable information between support and query sets to achieve predictions [25]. In a data-scarce scenario, it has been proven to improve the performance of few-shot learning over inductive solutions [28–30]. TPRN treated the sample relation of each support-query pair as a graph node, then resorted to the known relations between support samples to estimate the relational adjacency among the different support-query pairs [31]. DSN proposed an extension of existing dynamic classifiers by using subspaces and introduced a discriminative formulation to encourage maximum discrimination between subspaces during training, which avoids over-performing and boosts robustness against perturbations [32]. Huang et al. proposed PTN to revise the Poisson model tailored for few-shot problems by incorporating the query feature calibration and the Poisson MBO (Merriman-Bence-Osher) model to tackle the cross-class bias problems due to the data distribution drift between the support and query data [26]. EGNN exploited the edge labels rather than the node labels on the graph to exploit both intra-cluster similarity and inter-cluster dissimilarity to evolute an explicit clustering [33]. Unlike the above methods, in this paper, we adopt the transductive graph approach to construct the ETPN model. It leverages the related prior knowledge between support and query sets during the test phase and a novel fusion strategy to address the issue of high variance and over-fitting.

(2) Semi-supervised Few-shot Learning

Moreover, it is difficult to annotate samples in many fields, such as medicine, military, finance, etc. Thus, semi-supervised few-shot learning (SSFSL) approaches are proposed to leverage the extra unlabeled data to enhance the performance of few-shot learning. LTTL proposed a self-training model, which utilizes cherry-picking to search for valuable samples from pseudo-labeled data via a soft-weighting network [34]. PRWN proposed prototypical random walk networks to promote prototypical magnetization of the learning representation [35]. BR-ProtoNet exploited unlabeled data and constructed complementary constraints to learn a generalizable metric [36]. In this paper, we adopt transductive

inference to utilize unlabeled data and distractor classes irrelevant to the classification task to boost robustness against perturbations.

(3) Ensemble Few-Shot Learning

Ensemble learning is widely used in classifications to enhance the generalization ability and robustness of models. Therefore, many researchers have applied the ensemble framework to few-shot learning. The main idea is to adopt a combination approach to reduce the over-fitting problem and enhance the stability of the model. DIVERSITY investigated an ensemble approach for training multiple convolutional neural networks (CNNs). Each network predicts class probabilities, which are then integrated by a mean centroid classifier constructed for each network. Moreover, it introduced penalty terms allowing the networks to cooperate during training to guarantee the diversity of predictions [37]. EBDM divided the feature extraction network into shared and exclusive components. The shared component aims to share and reduce parameters in the lower layers, while the exclusive component is designed to be unique to each learner in the higher layers [38]. HGNN proposed a novel hybrid GNN of a prototype and instance to address overlapping classes and outlying samples, respectively [39].  $E^{3}BM$  introduced a Bayes model for each epoch, which leverages innovative hyperprior learners to learn task-specific hyperparameters and enhances model robustness [40]. However, the existing integration strategies mainly adopt a max-voting strategy without considering information uncertainty. Different from the above methods, we propose an improved D-S method to solve the above problem by preprocessing the data source; moreover, we improved the ensemble pruning method to perform a selective ensemble with better accuracy.

The contribution of our algorithm is summarized in Table 1, including transduction inference (trans\_inference), semi-supervised few-shot learning (SSFSL), ensemble, ensemble pruning, and information uncertainty (infor\_uncertainty).

Model	trans_inference	SSFSL	Ensemble	Ensemble Pruning	infor_uncertainty
TPRN [31]	$\checkmark$	×	×	×	×
DSN [32]	$\checkmark$	√ (w/D)	×	×	×
PTN [26]	$\checkmark$	√ (w/D)	×	×	×
EGNN [33]	$\checkmark$	$\checkmark$	×	×	×
LTTL [34]	×	$\checkmark$	×	×	×
PRWN [35]	×	√ (w/D)	×	×	×
BR – ProtoNet [36]	×	√ (w/D)	×	×	×
DIVERSITY [37]	×	×	$\checkmark$	×	×
EBDM [38]	×	×	$\checkmark$	×	×
HGNN [39]	$\checkmark$	×	$\checkmark$	×	×
$E^{3}BM$ [40]	×	×	$\checkmark$	×	×
IG – semiTPN [ <mark>41</mark> ]	$\checkmark$	√ (w/D)	×	×	×
He - ETPN(our)	$\checkmark$	√ (w/D)	$\checkmark$	$\checkmark$	$\checkmark$
Ho - ETPN(our)	$\checkmark$	√ (w/D)	$\checkmark$	$\checkmark$	$\checkmark$

Table 1. The key contributions. Where "w/D" means with distractors.

#### 3. Problem Definition

Given a label set  $C = \{c_j | j = 1, 2, ..., N\}$ ,  $c_j$  represents the label (i.e., a discrete value).  $S = \{s_i | i = 1, 2, ..., n\}$  denotes a sample set,  $s_i = (x_i, y_i)$  represents a sample,  $x_i$  is the attribute values set,  $||x_i||$  denotes number of dimensions, namely,  $x_i = (x_i^1, x_i^2, ..., x^{||x_i||})$ . If  $y_i \in C$ ,  $s_i$  represents labeled samples, otherwise  $s_i = (x_i, ...)$  represents unlabeled samples. Sample sets are divided into supervised represented  $S_{sup}$  and unsupervised sample sets represented by  $S_{uns}$ , thus,  $S = S_{sup} \cup S_{uns}$ , where  $S_{sup} = \{(x_i, y_i) | (x_i, y_i) \in S, y_i \in C\}$ ,  $S_{uns} = \{(x_i, ...) | (x_i, ...) \in S\}$ .

For  $S_{sup}$ , let  $X \longrightarrow C$  denote the process of predicting the labels by the classifier F for training samples, where  $X = \{x_i | (x_i, y_i) \in S_{sup}\}$ , namely  $\forall (x_i, y_i) \in S_{sup}$ ,  $\tilde{y}_i = F(x_i)$ . The

accuracy rate of a classifier *F* is defined as  $Acc(S_{sup}, F) = \frac{|\{(x_i, y_i)|(x_i, y_i) \in S_{sup}, \tilde{y_i} = F(x_i), \tilde{y_i} = y_i\}|}{|S_{sup}|}$ Supervised machine learning is the process of obtaining a classifier *F* from  $S_{sup}$ .

For *F* and  $S_{uns}$ , obtaining  $S_{uns \to sup}$  from  $S_{uns}$  is the process of adding annotations, which can be defined as  $S_{uns \to sup} = \{(x_i, \tilde{y_i}) | (x_{i, -}) \in S_{uns}, \tilde{y_i} = F(x_i)\}$ . For  $S = S_{sup} \cup$  $S_{uns}$  and  $X \longrightarrow C$ , where  $X = \{x_i | (x_i, y_i) \in S_{sup} or(x_i, \tilde{y_i}) \in S_{uns \to sup}\}$ , semi-supervised machine learning is the process of obtaining *F* from *S*.

For  $S = S_T \cup S_V$ , where  $S_T$  is the train set,  $S_V$  is validation set. F is learned from  $S_T$ . The validation is the process of calculating  $Acc(S_V, F)$ .

Few-shot learning constructing models generally adopt episodic training mode. According to the above notations, the episodic training (K-way, N-shot) is defined as follows: let the label set of the support set denote  $C_K \subseteq C$ ,  $|C_K| = K$ ,  $\forall C_K, \exists y \in C_K, S^N \subseteq S$ ,  $|S^N| = N, s \in S^N, s = (x, y)$ ; the support set is defined as  $T^{K \cdot N} = \{S_i^N | i = 1, 2, ..., K\}$ ,  $\exists y \in C_K$  and the query set is defined as  $Q^M \subseteq S$ ,  $|Q^M| = M, q \in Q^M, \forall S_i^N \in T^{K \cdot N}, q \notin S_i^N$ .

For  $S_{uns \rightarrow sup} = \{(x_i, \tilde{y}_i) | (x_i, \_) \in S_{uns}, \tilde{y}_i = F(x_i)\}$ , the interference sets consist of the distractor classes irrelevant to the target tasks [42,43]. It is added in the support and query sets to boost model robustness.

## 4. Methodology

In this paper, we propose the ensemble transductive propagation network (ETPN). The whole framework of the ensemble model is shown in Figure 1. For ETPN, we propose Ho-ETPN and He-ETPN models according to different ensemble approaches. Additionally, we incorporate a preset weight coefficient and compute iterative inferences during transductive propagation learning. Moreover, an improved D-S evidence fusion strategy is proposed for comprehensive evaluation. Meanwhile, we improve the ensemble pruning method to screen individual learners of higher accuracy to conduct fusing.



Figure 1. The overall framework diagram of the model.

There are several important parts in our ETPN model (as shown in Figure 2), including the framework of Ho-ETPN and He-ETPN, constructing KNN-Graphs using the improved Gaussian kernel [41], transductive propagation learning, and evidence fusion strategy. Next, we introduce the single model framework of IG-semiTPN simply, then introduce other parts of our ensemble model in detail.



Figure 2. The overall flow chart of the ensemble model.

## 4.1. IG-semiTPN Model

We propose our ensemble semi-supervised graph network based on the individual learner framework of IG-semiTPN [41] to utilize the information shared between support and query datasets. The framework of IG-semiTPN is shown in Figure 3.

Firstly, it employs  $f_{\varphi}(\cdot)$  to extract features of the input  $x_i$  and  $x_j$  ( $\varphi$  indicates a parameter of the network). Then, the graph construction module  $g_{\varphi}$  (as shown in Figure 4) is utilized to learn  $\sigma_i$  for every instance. Next, an improved Gaussian kernel  $w_{ij}(f_{\varphi}(x_i), f_{\varphi}(x_j), \sigma_i, \sigma_j)$  is proposed to calculate the edge weight for constructing the k-nearest neighbor graph. Finally, the label propagation method is adopted to achieve transductive propagation learning.



Figure 3. The framework of the IG-semiTPN model.



**Figure 4.**  $g_{\phi}$  construction of model.

#### 4.2. ETPN Model

Ensemble learning aims to enhance the generalization ability and stability of individual learners. The homogeneous framework employs a single base learning algorithm, i.e., learners of the same type but with multiple different sample inputs, leading to homogeneous ensembles (shown in Figure 5). The heterogeneous model utilizes multiple learning algorithms, i.e., learners of different types, leading to heterogeneous ensembles (shown in Figure 6).



Figure 5. The process of homogeneous ensembles.



Figure 6. The process of heterogeneous ensembles.

## 4.2.1. The Ho-ETPN Model

In this section, we propose a homogeneous ensemble few-shot learning model (Ho-ETPN, shown in Figure 7). The Ho-ETPN model generates multiple results (evidence) by randomly selecting different support sets (i.e.,  $support_i^n$ , being the same categories but different samples) in every episodic training. In contrast, the query set is selected only once in each episodic training. It generates multiple results by the same individual learner introduced in the last sector then integrates them via the the evidence fusion strategy proposed in this paper to accomplish predictions.



Figure 7. The framework of the Ho-ETPN model.

#### 4.2.2. The He-ETPN Model

In this section, we propose a heterogeneous ensemble few-shot learning model (He-ETPN). The He-ETPN model generates multiple results (evidence) via multiple learners in every episodic training, and then integrates them via the evidence fusion strategy proposed in this paper to accomplish predictions. The model framework of base learners has been introduced in the last section. The He-ETPN model (shown in Figure 8) generates multiple results by constructing diverse KNN-graphs using different models that are the different initializations of  $f_{\varphi}(\cdot)$  and  $g_{\phi}$ , with different value settings of  $\gamma$  and m in an improved Gaussian kernel.



Figure 8. The framework of the He-ETPN model.

## 4.2.3. Construct KNN-Graphs

For dataset  $S = S_{sup} \cup S_{uns}$ , during the construction of KNN-graphs, let  $s \in S$  represent the graph vertex to build the undirected graphs of labeled and unlabeled samples. We use the edge weights to measure the similarity between samples, the greater the weight of the edge, the greater the similarity between the two samples. Due to the improved Gaussian kernel [41] the nuclear truncation effect is alleviated by adding displacement parameters and corrections and learning a  $\sigma$  parameter suitable for each sample in the process of constructing the graph. Therefore, we utilize the improved Gaussian kernel to calculate the edge weight to construct more accurate KNN graphs for the transductive propagation ensemble network. The improved Gaussian kernel is defined as follows:

$$w_{ij} = exp(\frac{1}{md(\frac{f_{\varphi}(x_i)}{\sigma_i}, \frac{f_{\varphi}(x_j)}{\sigma_i}, m) + \gamma^2} + \lambda), i, j = 1, \dots, n,$$
(1)

where  $f_{\varphi}(\cdot)$  refers to the feature map,  $\varphi$  indicates a parameter of the network,  $\sigma$  is the variable bandwidth (length scale parameter) of the kernel function learned by  $g_{\varphi}$ ,  $\gamma$  is the displacement parameters,  $\lambda$  is the fine-tuning variable.  $d(\cdot)$  is the Euclidean distance.

## 4.3. Transductive Propagation Learning

Transductive propagation learning aims to predict unlabeled data from locally labeled training samples. It takes the support and query set as graph nodes, then makes joint predictions using the nearest neighbor relationship between the support and query sets (as shown in Figure 9), which can supplement the deficiency of label information through unlabeled data. Due to its low complexity and good classification effect, the label propagation algorithm is adopted to perform the transfer of label information between graph nodes. The process of predictions for the query set  $Q^M$  using label propagation is defined as follows:



Figure 9. Transductive propagation algorithm based on K-nearest neighbor graph.

(1) Suppose  $F = [Y_l, Y_u]$  is an annotation matrix with  $(K \cdot N + M) \times K$  – *dimension*.  $Y_l$  denotes the support set sample label matrix, and  $Y_u$  denotes the query set sample matrix Let Y is the initial annotation matrix  $Y \in F$ ,  $Y_{ic}$  represents the membership degree to the

c - th column category of i - th node  $Y_i$ . If  $Y_{ic} = 1$ , which is mean the node  $Y_i$  belonging to the category c, else  $Y_{ic} = 0$ , that is,

$$Y_{ic} = \begin{cases} 1, \ Y_i = c \\ 0, \ otherwise \end{cases}$$
(2)

(2) Given the initial label matrix  $Y_l$ , the query set labels are iteratively predicted according to the KNN-graphs. Let  $\alpha \in (0, 1)$  denote the pre-set weight coefficient to control the amount of propagated information. The transductive propagation learning iteratively inferences as follows:

 $\lim_{x\to\infty}T^t=\begin{pmatrix}I&0\\\alpha\cdot T_{ul}&\alpha\cdot T_{uu}\end{pmatrix}\times\begin{pmatrix}I&0\\\alpha\cdot T_{ul}&\alpha\cdot T_{uu}\end{pmatrix}\times\begin{pmatrix}I&0\\\alpha\cdot T_{ul}&\alpha\cdot T_{uu}\end{pmatrix}\times\ldots$ 

$$=\begin{pmatrix}I \cdot I + 0 + 0 + \dots & 0 + 0 + 0 + \dots \\ (\alpha \cdot T_{ul}) + (\alpha \cdot T_{uu}) \cdot (\alpha \cdot T_{uu}) \cdot (\alpha \cdot T_{uu})^2 + \dots & (\alpha \cdot T_{uu}) \cdot (\alpha \cdot T_{uu}) \cdot (\alpha \cdot T_{uu}) \dots\end{pmatrix}$$

$$=\begin{pmatrix} I \cdot I + 0 + 0 + \dots & 0 + 0 + 0 + \dots \\ (I + (\alpha \cdot T_{uu}) + (\alpha \cdot T_{uu})^2 + \dots) \cdot (\alpha \cdot T_{ul}) & (\alpha \cdot T_{uu}) \cdot (\alpha \cdot T_{uu}) \cdot (\alpha \cdot T_{uu}) \dots \end{pmatrix}$$
(3)

$$= \begin{pmatrix} I & 0 \\ (\sum_{t=0}^{\infty} (\alpha \cdot T_{uu})^t) \cdot (\alpha \cdot T_{ul}) & (\alpha \cdot T_{uu})^{\infty} \end{pmatrix}$$

$$= \begin{pmatrix} I & 0\\ (I - \alpha T_{uu})^{-1} \cdot (\alpha \cdot T_{ul}) & 0 \end{pmatrix}$$

$$F_{t+1} = \lim_{x \to \infty} F_t = \lim_{t \to \infty} T^t \begin{pmatrix} (1 - \alpha)Y_l\\ 0 \end{pmatrix} = \begin{pmatrix} I & 0\\ (I - \alpha T_{uu})^{-1} \cdot (\alpha \cdot T_{ul}) & 0 \end{pmatrix} \begin{pmatrix} (1 - \alpha)Y_l\\ 0 \end{pmatrix}, \quad (4)$$

where  $T = D^{-1/2}WD^{-1/2}$ , *T* represents the normalized Laplace matrix.  $T_{ll}$  denotes the identity matrix;  $T_{lu}$  denotes the zero matrix;  $T_{ul} \subset T$ ;  $T_{uu} \subset T$ .  $W = [w_{ij}]$ ,  $W \in R^{(K \cdot N + M) \times (K \cdot N + M)}$  for all instances in  $S = S_{sup} \cup S_{uns}$ .

We keep the k - max values in each row of W calculating by Equation (1) to construct KNN-graphs. Then, the normalized graph Laplacian is applied [44] on W; D is the diagonal matrix,  $D = diag([d_{ii} = \sum_{j=1}^{K \cdot N+M} w_{ij}])$ . While  $\alpha$  is bigger, the results tend to favor label propagation items  $T_{uu}$  and  $T_{ul}$ , else, results prefer the original annotated items  $Y_l$ . The final prediction results  $F^*(F^* = [f_{i\eta}])$  are obtained through multiple iterations, as is shown in Equation (5).

$$F^* = \lim_{x \to \infty} F_t = (I - \alpha T_{uu})^{-1} (\alpha \cdot T_{ul}) \cdot (1 - \alpha) Y_l.$$
(5)

# 4.4. Ensemble Pruning

The error-ambiguity decomposition [45] can show that the success of ensemble learning depends on a good trade off between the individual performance and diversity, which is defined as follows:

$$err(H) = \overline{err}(h) - \sqrt{ambi(h)},$$
(6)

where  $\overline{err}(h) = \sum_{i=1}^{N} \omega_i \cdot err(h_i)$  denotes the average error of the individual learners, and  $\overline{err}(h) = \sum_{i=1}^{N} \omega_i \cdot ambi(h_i)$  denotes the weighted average of ambiguities. *h* denotes the individual learner.

The  $\overline{err}(h)$  depends on the generalization ability of individual learners; the ambi(h) depends on the ensemble diversity. Since the  $\overline{ambi}(h)$  is always positive, obviously, the error of the ensemble will never be larger than the average error of the individual learners. More importantly, Equation (6) shows that the more accurate and the more diverse the individual learners, the better the ensemble. Based on this, we propose an improved ensemble pruning approach to select more accurate learners to participate in the integration.

Ensemble pruning is to associate an individual learner with a weight that could characterize the goodness of including the individual learner in the final ensemble. RSE is a regularized selective ensemble algorithm; it adopted the L1 norm for feature selection to obtain sparse weights [46]. However, the sample space of few-shot learning is more sparse. To enhance data utilization and ensure that samples far away from decision boundaries still contribute to model training, we employ the L2 norm to obtain weights as small as possible but not zero. In addition, this makes the model more stable to small changes of the input and improves the robustness of the model. Moreover, to be suitable for few-shot learning, we redefine the improved ensemble pruning algorithm.

Given *n* individual learners for He-ETPN or Ho-ETPN, let  $\emptyset = [\emptyset_1, \dots, \emptyset_n]$  denote the n-dimensional weight vector of *n* individual learners, where small elements in the weight vector suggest that the corresponding individual learner from He-ETPN or Ho-ETPN should be excluded during the process of fusion.

$$\Theta(\omega) = \lambda \Lambda(\omega) + \Omega(\omega), \tag{7}$$

where  $\Lambda(\omega)$  is the empirical loss,  $\Omega(\omega)$  is the graph Laplacian regularization term to measure the misclassification, and  $\lambda$  is a regularization parameter which trades off the minimization of  $\Lambda(\omega)$  and  $\Omega(\omega)$ .

By introducing slack variables  $\eta$  and minimizing the regularized risk function to determine the weight vector, Equation (7) is redefined as follows:

$$\min_{\omega} \omega P^{\top} T P \omega^{\top} + \lambda (\eta_1^2 + \ldots + \eta_m^2)$$
  
s.t.  $y_i p_i \omega^{\top} + \eta_i^2 \ge 1, (\forall i = 1, \ldots, m)$   
 $\omega_1 + \ldots + \omega_m = 1, \omega \ge 0,$  (8)

where *P* denotes the prediction matrix of all individual learners on all support set instances,  $p_i = (\max(F_1^*), \dots, \max(F_n^*))$  denotes the predictions of the individual learner on  $x_i$ , and *T* represents the normalized Laplace matrix.  $y_i$  denotes the sample label of  $x_i$ .  $\eta = (\eta_1, \dots, \eta_m)^\top$  denotes the slack variables.

$$\varpi_i = \begin{cases}
1, \ top - n \\
0, \ otherwise
\end{cases}$$
(9)

$$F_i^* = \mathcal{O}_i \cdot F_i^*, i = 1, \dots, n, \tag{10}$$

where top - n is to select the top n best individual learners for the pruned ensemble.  $F_i^*$  denotes a piece of evidence, if the  $\omega_i = 0$  denotes that the  $F_i^*$  does not participate in the fusion of the results.

The complexity of the pruning approach is  $O(n^3)$ . Equation (8) is a standard QP problem that can be efficiently solved by existing optimization packages. It is more suitable for small-scale datasets, especially few-shot learning.

#### 4.5. Evidence Fusion Strategy

In this paper, we propose an improved D-S evidence fusion method to assemble the multiple pieces of evidence generated by the ensemble solutions of the Ho-ETPN and He-ETPN. Compared with the averaging and voting methods, the improved D-S evidence fusion method can enhance the stability of ensemble results and alleviate the problem of the "Zadeh paradox" to a certain extent. The D-S evidence theory was first proposed by Dempster [47,48]. Combining multiple information sources is an effective method of uncertainty reasoning. The research indicates that the synthetic consequence of conventional combination rules of Dempster is frequently contrary to the reality in the practical applications [49,50]. Two major approaches are proposed to improve the accuracy of synthetic results—one is to amend the composition rules; the second is to change the original evidence resources. In this paper, we focus on the latter. Next, we concretely introduce the process of the improved D-S evidence fusion method.

(1) Conflict Matrix

The Bhattacharyya distance [51] is utilized to construct the conflict matrix between evidence. According to the intension of Bhattacharyya distance, the formula is redefined as follows:

**Definition 1** (Bhattacharyya Distance). For probability distributions  $F_i^*$  and  $F_j^*$  over the same domain, the Bhattacharyya distance is defined as:

$$Dis_{BC}(F_i^*, F_j^*) = \sum_{\eta=1}^{K} \sqrt{F_i^*(f_{i\eta})F_j^*(f_{j\eta}))}$$
(11)

$$fc_{ij} = -ln(Dis_{BC}(F_i^*, F_j^*)),$$
(12)

where  $Dis_{BC}(F_i^*, F_j^*)$  is the Bhattacharyya coefficient for discrete probability distributions. Let  $n = K \cdot N + M$  denote the number of pieces of evidence. Each piece of normalized evidence is denoted by  $F_i^* = (f_{i1}, f_{i2}, \ldots, f_{iK})$ .  $F_i^*$  and  $F_j^*(1 \le i, j \le n)$  represent two pieces of evidence. The K denotes the number of classes in each support set,  $k_i \in K$ . Then, the normalization conflict matrix is defined as:

$$Matrix_{conflict} = \begin{bmatrix} 0 & fc_{12} & \cdots & fc_{1j} & \cdots & fc_{1n} \\ fc_{21} & 0 & \cdots & fc_{2j} & \cdots & fc_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ fc_{j1} & fc_{j2} & \cdots & 0 & \cdots & fc_{jn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ fc_{n1} & fc_{n2} & \cdots & fc_{nj} & \cdots & 0 \end{bmatrix}.$$
(13)

## (2) Support Degree

Evidence support degree indicates the support degree of evidence that is supported by other evidence. The higher the similarity with other evidence, the higher the support degree it is, and vice versa. According to  $Matrix_{conflit}$  the following formula is utilized to calculate the similarity degree between  $F_i^*$  and  $F_i^*$ .

$$sd_{ij} = 1 - fc_{ij}, \quad i, j = 1, 2, \dots, n.$$
 (14)

As a result, we can obtain the following similarity matrix of all evidence:

$$Matrix_{similarity} = \begin{vmatrix} 1 & sd_{12} & \cdots & sd_{1j} & \cdots & sd_{1n} \\ sd_{21} & 1 & \cdots & sd_{2j} & \cdots & sd_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ sd_{j1} & sd_{j2} & \cdots & 1 & \cdots & sd_{jn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ sd_{n1} & sd_{n2} & \cdots & sd_{nj} & \cdots & 1 \end{vmatrix}.$$
(15)

And then, the support degree of each evident is calculated as:

$$Sup(F_i^*) = \sum_{j=1, j \neq i}^n sd_{ij}.$$
(16)

(3) Evident Weight

Credibility degree indicates the credibility of an evidence. It can be calculated by following formula.

$$CR(F_i^*) = \frac{Sup(F_i^*)}{\sum_{j=1}^n Sup(F_j^*)}.$$
(17)

Information entropy can be utilized to measure the informative quantity of evidence in the information fusion process. Integrated with D-S evidence theory, given a piece of evidence  $F_i^* = (f_{i1}f_{i2}, \ldots, f_{iK})$ , and  $\sum_{\eta=1}^{K} f_{i\eta} = 1$ . The information quantity of the *ith* piece of evidence is defined as:

$$Info_e(F_i^*) = -\sum_{\eta=1}^K f_{i\eta} log f_{i\eta}.$$
(18)

For information entropy, the larger the uncertainty, the smaller its weight. On the other hand, the smaller the information entropy, the larger its weight. The method mentioned above can be used to reduce the weight ratio of the evidence with higher indeterminacy in the fusion process. Therefore, the weight of each evidence is defined as:

$$weight(F_i^*) = \frac{CR(F_i^*)}{Normalization(Info_e(F_i^*))}, \ 1 \le i \le n$$
(19)

$$F_i^* = weight(F_i^*) \times F_i^*.$$
<sup>(20)</sup>

(4) Evidence Combination Rule

Suppose that the feature subsets generated in the previous chapter are independent. The D-S evidence theory improved in this paper allows the fusion of information coming from different feature subsets. Therefore, the evidence combination rule is utilized to combine different weighted feature subsets in a manner that is both accurate and robustness.

For  $F_i^*$  (i = 1, 2, ..., n),  $\forall k \in K$ , the combination rule is redefined as:

$$F_{fusion}^{*} = (F_{1}^{*} \ \uplus \ F_{2}^{*} \ \uplus \cdots \uplus \ F_{n}^{*})(k) = \frac{1}{Q} \prod_{i=1}^{n} F_{i}^{*}(k), \tag{21}$$

where *Q* means the conflict between different pieces of evidence, is given by:

$$Q = \sum_{i=1}^{K} \prod_{j=1}^{n} F_j^*(k_i).$$
(22)

#### 4.6. Loss Generation

In this paper, we adopt cross-entropy loss to calculate the similarity between predictive values and true values.

(1) We adopt the softmax function to transform the  $F_{fusion}^*$  of the ETPN model to probability, which is defined as follows:

$$P(\widetilde{y}_i = j | x_i) = \frac{exp(f_{i\eta})}{\sum_{j=1}^{N} exp(f_{i\eta})},$$
(23)

where  $\tilde{y}_i$  is the final prediction value of i - th samples in query.  $f_{i\eta}$  is the component of the prediction values in label propagation.

(2) We calculate the loss by the cross entropy loss:

$$J(\varphi, \phi) = \sum_{i=1}^{T} \sum_{j=1}^{N} -\mathbb{I}(y_i = j) log(P(\widetilde{y}_i = j | x_i)),$$
(24)

where  $y_i$  is the true value of the instance.  $\mathbb{I}(b)$  is the indicator function. If *b* is right, and  $\mathbb{I}(b) = 1$ , else  $\mathbb{I}(b) = 0$ .

## 5. Experiments

In this section, to validate the performance of models, we contrast our proposal against state-of-the-art techniques on miniImageNet and tieredImageNet datasets. In addition, we set a supervised experiment including the ensemble model Ho-ETPN and He-ETPN, and a semi-supervised experiment including the setting of distractor classes. These approaches are particularly divided into optimization-based (MAML [22]), ensemble-based (EBDM-Euc [38], HGNN [39], E<sup>3</sup>BM+MAML [40]), graph-based (TPN [25], EPNet [27], TPRN [31], DSN [32], EGNN [33], PRWN [35], GNN [52], BGNN\* [53], DPGN\* [54]), and metric-based (MatchingNet [8], Proto Net [9], TADAM [13], BR-ProtoNet [36], SSFormers [55], CGRN [56], HMRN [57]) approaches. Moreover, we conduct 5-way 1-shot and 5-shot experiments, which are standard few-shot learning settings.

#### 5.1. Datasets

**miniImageNet** [8]. A subset of the ImageNet datasets [58] consists of 60,000 images. Each image is of size  $84 \times 84$ , and classes with 600 samples per class are divided into 64, 16, and 20 for meta-training, meta-validation, and meta-testing, respectively. We use the miniImageNet for semi-supervised classification with 40% of labeled data.

**tieredImageNet** [41]. A more challenging subset derived from ImageNet datasets, its class subsets are chosen from supersets of the wordnet hierarchy. The top hierarchy has 34 super-classes, which are split into 20 different categories (351 classes) for training, six different categories (97 classes) for validation, and eight different categories (160 classes ) for testing. We follow the implementation of 4-convolutional layer (*Conv* – 4) backbones and the image size of 84  $\times$  84 as on miniImageNet. Moreover, the tieredImageNet is used for semi-supervised classification with 10% of labeled data.

## 5.2. Implementation Details

Following the Matching Networks [8], we also adopt the episodic training procedure. Moreover, we used a common feature extractor, which is a *Conv* – 4 as implemented in [8] during the entire comparision experiments for standard few-shot classification. It makes up four convolutional blocks where each block begins with a 2D convolutional layer with a 3 × 3 kernel and a filter size of 64. Each convolutional layer is followed by a batch-normalization layer [43], a ReLU nonlinearity, and a 2 × 2 max-pooling layer. Moreover,  $g_{\phi}$  utilized to learn  $\sigma_i$  for every instances, consists of two convolutional blocks (64 and 1 filters) and two fully-connected layers (8 and 1 neurons) similar to TPN [25]. The convolutional blocks are made up of four convolutional blocks and each block begins with a 2D convolutional layer with a 3 × 3 kernel and filter size of 64. Each convolutional layer is followed by a batch-normalization layer [43], a ReLU nonlinearity and a 2 × 2 max-pooling layer. In the experiments, we follow a general practice to evaluate the model with N-way K-shot and 15 query images; the value of  $\lambda$  is set to 0.75. And we use Adam optimizer [59] with an initial learning rate of 0.001, we use the validation set to select the training episodes with the best accuracy, and run the training process until the validation loss reaches a plateau.

In addition, we utilize the improved Gaussian kernel proposed in the single model framework IG-semiTPN to construct the KNN graphs. IG-semiTPN experiments showed the superior effects of the improved Gaussian kernel function. It also indicated that the optimal models have relations with the value of  $\gamma$  [60,61]. Therefore, ETPN utilizes the parameter settings of the improved Gaussian kernel of the IG-semiTPN to perform supervised and semi-supervised experiments. Specifically, Ho-ETPN adopts the Minkowski distance with  $\gamma$  being 3 and m being 3 or the Minkowski distance with m being 2 and  $\lambda$  is 0.75. In addition, there are three learners in our He-ETPN ensemble models; learner 1 adopts a Minkowski distance with  $\gamma$  being 0.2 and m being 2; learner 3 adopts a Minkowski distance with m being 2, and  $\lambda$  is 0.75.

# 5.3. Supervised Experiment ETPN Experiment

In our experiments, we compare ensemble model ETPN with other classic and advanced algorithms in four categories, including graph-based (TPN [25], EPNet [27], TPRN [31], DSN [32], EGNN [33], PRWN [35], GNN [52], BGNN\* [53], DPGN\* [54]), metric-based (MatchingNet [8], Proto Net [9], TADAM [13], BR-ProtoNet [36], SSFormers [55], CGRN [56], HMRN [57]), optimization-based (MAML [22]) and ensemble-based (EBDM-Euc [38], HGNN [39], E<sup>3</sup>BM+MAML [40]) approaches. The performance of the proposed ETPN and state-of-the-art models in the 5-way 5-shot/1-shot accuracy on the miniImageNet and tieredImageNet datasets are summarized in Tables 2 and 3, and "\*" in the table indicates results re-implemented in HGNN [39] for a fair comparison. Our proposed ETPN outperforms few-shot models by large margins, indicating that the proposed ensemble model effectively assists few-shot recognition. Specifically, we can obtain the following observations:

(1) **Comparison with the latest model.** Ho-ETPN is 8.32% higher than SSFormers in 5-shot on miniImageNet and 5.58% higher than HMRN in 5-shot on tieredImageNet; Ho-ETPN is 7.86% higher than SSFormers in 1-shot on miniImageNet and 9.59% higher than HMRN in 1-shot on tieredImageNet. It confirms that our model has a good ability for classification discrimination, which benefits from our ensemble model and D-S evidence fusion strategy based on improved ensemble pruning.

(2) **Comparison with the state-of-the-art.** Under the 5-way-5-shot setting, the ETPN classification accuracies are 78.87% vs. 78.57% for the transductive learning model TPRN, 80.28% vs. 80.0% ensemble model BR-ProtoNet on miniImageNet and tieredImageNet, respectively. It is 0.3% higher than the transductive learning model TPRN in 5-shot on miniImageNet and 0.28% higher than the ensemble model BR-ProtoNet in 5-shot on tiered-ImageNet. Under the 5-way-1-shot setting, the ETPN classification accuracies are 63.06% vs. 57.84% for the transductive learning model TPRN, 67.57% vs. 62.7% for the ensemble model BR-ProtoNet on miniImageNet and tieredImageNet, respectively. It is 5.22% higher than the transductive model TPRN in 1-shot on miniImageNet and 4.87% higher than the ensemble model BR-ProtoNet in 1-shot on tieredImageNet. ETPN outperforms state-of-the-art few-shot models by large margins, especially under the 5-way-1-shot setting. This indicates that our ensemble model and improved evidence fusion strategy are effective, particularly for scenarios with a small sample size, which can increase the performance by enhancing the stability of the model.

Methods	т	γ	λ	5way5shot	5way1shot
MatchingNet [8]				55.31%	43.56%
Proto Net [9]				68.20%	49.42%
TADAM [13]				76.7%	58.5%
BR-ProtoNet [36]				74.5%	58.4%
Relation Network [62]				65.32%	50.44%
MAML [22]				63.11%	48.70%
PRWN [35]				67.82%	50.89%
TPN [25]				67.79%	53.42%
EPNet [27]				72.95%	59.32%
TPRN [31]				78.57%	57.84%
DSN [32]				68.99 %	51.78%
EGNN [33]				66.85%	59.63%
GNN [52]				66.41%	50.33%
BGNN* [53]				67.35	52.35%
DPGN* [54]				65.34	53.22%
EBDM-Euc (2heads) [38]				68.30%	49.96%
EBDM-Euc (3heads) [38]				69.14%	50.49%
EBDM-Euc (5heads) [38]				69.64%	52.53%
EBDM-DD (2heads) [38]				67.99%	51.42%
EBDM-DD (3heads) [38]				68.74%	52.56%
EBDM-DD (5heads) [38]				70.17%	53.08%
HGNN [39]				72.48%	55.63%
$E^{3}BM + MAML$ [40]				65.1%	53.2%
SSFormers [55]				70.55%	55.2%
CGRN [56]				64.13%	50.85%
He-ETPN (our)			0.75	72.94%	59.87%
Ho-ETPN (our)	3	3	0.75	77.13%	<b>63.06</b> %
Ho-ETPN (our)	2		0.75	78.87%	62.33%

 Table 2. Few-shot classification accuracies on the miniImageNet dataset are cited.

 Table 3. Few-shot classification accuracies on the tieredImageNet dataset are cited.

Methods	т	$\gamma$	$\lambda$	5way5shot	5way1shot
Proto Net [9]				72.69%	53.31%
MAML [22]				70.30%	51.67%
PRWN [35]				70.52%	54.87%
BR-ProtoNet [36]				80.0%	62.7%
Relation Network [62]				71.32%	54.48%
TPN [25]				71.2%	56.17%
EGNN [33]				70.96%	
TPRN [31]				79.66%	59.26%
HGNN [39]				72.82%	56.05%
BGNN* [53]				65.27%	49.41%
DPGN* [54]				69.86%	53.99%
EPNet [27]				73.91%	59.97%
EBDM-Euc (3 heads) [38]				72.24%	51.22%
EBDM-Euc (1-st head) [38]				71.07%	50.04%
EBDM-Euc (2-nd head) [38]				71.28%	50.29%
EBDM-Euc (3-rd head) [38]				70.84%	50.52%
$E^{3}BM + MAML$ [40]				70.2%	52.1%
SSFormers [55]				73.72%	55.54%
CGRN [56]				71.34%	55.07%
HMRN [57]				74.70%	57.98%
He-ETPN (our)			0.75	74.08%	62.75%
Ho-ETPN (our)	3	3	0.75	80.28%	66.24%
Ho-ETPN (our)	2		0.75	80.18%	67.57%

(3) Compare with individual learner IG-semiTPN. In this section, we compare our ensemble model ETPN with our single supervised model IG-semiTPN [41]; this is to show that the homogeneous strategy, heterogeneous strategy, and improved D-S evidence fusion strategy based on improved ensemble pruning facilitate the model performance. For the fairness of the experiment, we ensure other settings are the same, only changing the ensemble strategy and the parameter settings in the improved Gaussian kernel to perform the ablation experiment. The comparison results are shown in Figures 10 and 11. Under the 5-way 5-shot setting, the classification accuracies of Ho-ETPN and IG-semiTPN are 78.87% vs. 69.31% on miniImageNet, and 80.28% vs. 73.21% on tieredImageNet, respectively. Ho-ETPN is 9.56% and 7.07% higher than IG-semiTPN in 5-shot on miniImageNet and tieredImageNet, respectively. Under the 5-way 1-shot setting, the classification accuracies of Ho-ETPN and IG-semiTPN are 63.06% vs. 54.03% on miniImageNet, and 67.57% vs. 57.35% on tieredImageNet, respectively. Ho-ETPN is 9.03% and 10.22% higher than IG-semiTPN in 1-shot on miniImageNet and tieredImageNet, respectively. In addition, under the 5-way 5-shot setting, the classification accuracies of He-ETPN and IG-semiTPN are 72.94% vs. 69.31% on miniImageNet, and 74.08% vs. 73.21% on tieredImageNet, respectively. He-ETPN is 3.63% and 0.87% higher than IG-semiTPN in 5-shot on miniImageNet and tieredImageNet, respectively. Under the 5-way 1-shot setting, the classification accuracies of He-ETPN and IG-semiTPN are 59.87% vs. 54.03% on miniImageNet, and 62.75% vs. 57.35% on tieredImageNet, respectively. He-ETPN is 5.84% and 5.4% higher than IGsemiTPN in 1-shot on miniImageNet and tieredImageNet, respectively. The results indicate the effectiveness of the proposed ensemble solutions, which achieve a state-of-the-art performance compared to the single model IG-semiTPN, especially in 1-shot. Moreover, the Ho-ETPN is superior to the He-ETPN, which is related to the problem of multiple learner selection. In our paper, we only select different parameter settings of  $f_{\varphi}$ ,  $g_{\phi}$  and an improved Gaussian kernel.





Figure 10. Comparison of the IG-semiTPN and ETPN (He-ETPN and Ho-ETPN) on miniImageNet.

Figure 11. Comparison of the IG-semiTPN and ETPN (He-ETPN and Ho-ETPN) on tieredImageNet.

#### 5.4. Semi-Supervised Experiment

Since labeled data are scarce and their collection is expensive, in this section, we leverage the extra unlabeled data to improve the performance of few-shot classifiers. Our model was trained on miniImageNet and tieredImageNet with 40% and 10% of labeled data, respectively. What is more, another key challenge is that the distractor classes, being an unlabeled set that is irrelevant to the classification task, are introduced to boost robustness

against perturbations. We follow the settings in papers [41,63]. Our models outperforms inference (TADAM-semi [13], BR-ProtoNet [36] and PN+Semi [41]) and transduction (TPN-semi [25], Semi-EPNet [27], Semi DSN [32], Semi-EGNN [33]and PRWN-semi [35]) semi-supervised few-shot models by large margins.

(1) **Comparison with the state-of-the-art.** In order to ensure the effectiveness of the semi-supervised experiment, every category in the datasets was divided into labeled datasets and unlabeled datasets without intersection [39]. In this paper, we utilize the label propagation algorithm to perform the annotation for unlabeled data, which is different from traditional inductive reasoning semi-supervised approaches. As is shown in Tables 4 and 5, it can be observed that the classification results of all semi-supervised few-shot models are degraded due to the distractor classes. However, even with the distractor class represented as w/D in the table, the ensemble semi-supervised model semi-HoTPN achieves the highest performance among the compared methods, especially in the scenario of 1-shot, which indicates the robustness of the proposed semi-HoTPN in dealing with distracted unlabeled data. In addition, this indicates that the proposed D-S evidence fusion strategy based on improved ensemble pruning, transductive propagation learning and homogeneous ensemble semi-supervised model semi-HoTPN effectively assists few-shot recognition.

Table 4. Semi-supervised comparison on the miniImageNet dataset.

Models	т	γ	λ	5way5shot	5way1shot	5way5shot (w/D)	5way1shot (w/D)
PN + Semi [41]				63.77%	49.98%	62.62%	47.42%
Soft k-Means [41]				64.59%	50.09%	63.55%	48.70%
Soft k-Means + Cluster [41]				63.08%	49.03%	61.27%	48.86%
Masked Soft k-Means [41]				64.39%	50.41%	62.96%	49.04%
TADAM-semi [13]						68.92%	54.81%
BR-ProtoNet [36]				73.1%	57.4%	72.4%	55.9%
Semi-EPNet [27]				67.08%			
Semi DSN [32]						67.12%	51.01%
Semi-EGNN [33]				64.32%			
PRWN-semi [35]				69.65%	56.65%	67.45%	53.61%
TPN-semi [25]				64.95%	50.43%	64.95%	50.43%
semi-HoETPN (our)	3	3	0.75	73.87%	60.57%	72.64%	59.34%
semi-HoETPN (our)	2		0.75	73.74%	61.31%	73.24%	59.28%

Table 5. Semi-supervised comparison on the tieredImageNet dataset.

Models	т	γ	λ	5way5shot	5way1shot	5way5shot (w/D)	5way1shot (w/D)
PN + Semi [41]				69.37%	50.74%	67.46%	48.67%
Soft k-Means [41]				70.25%	51.52%	68.32%	49.88%
Soft k-Means + Cluster [41]				69.42%	51.85%	67.56%	51.36%
Masked Soft k-Means [41]				69.88%	52.39%	69.08%	51.38%
BR-ProtoNet [36]				79.1%	61.8%	77.4%	60.1%
TPN-semi [25]				71.01%	55.74%	69.93%	53.45%
Semi DSN [32]						70.15%	53.89%
PRWN+Semi [35]				71.06%	59.17%	69.58%	56.59%
semi-HoETPN (our)	3	3	0.75	77.87%	63.87%	77.17%	63.16%
semi-HoETPN (our)	2		0.75	78.94%	65.21%	78.45%	64.80%

(2) **Compare with individual learner IG-semiTPN.** In this section, we show that the semi-supervised homogeneous ensemble model and improved D-S evidence fusion strategy based on improved ensemble pruning facilitate the model performance. For the fairness of the experiment, we compare semi-HoETPN with IG-semiTPN and other settings are the same. The comparison results are shown in Figures 12 and 13. We compare ensemble semi-supervised model semi-HoETPN with single semi-supervised model IG-semiTPN; under the 5-way-5-shot setting, the classification accuracies of semi-HoTPN

and IG-semiTPN are 73.87% vs. 67.24% on miniImageNet, and 78.94% vs. 72.32% on tieredImageNet, respectively. Under the 5-way-1-shot setting, the classification accuracies of semi-HoTPN and IG-semiTPN are 61.31% and 53.48% on miniImageNet, and 65.21% and 57.28% on tieredImageNet, respectively. With the distractor class experiments, under the 5-way-5-shot setting, the classification accuracies of semi-HoTPN and IG-semiTPN are 73.24% vs. 66.8% on miniImageNet, and 78.45% vs. 70.08% on tieredImageNet, respectively; under the 5-way-1-shot setting, the classification accuracies of semi-HoTPN and IG-semiTPN are 73.24% vs. 66.8% on miniImageNet, and 78.45% vs. 70.08% on tieredImageNet, respectively; under the 5-way-1-shot setting, the classification accuracies of semi-HoTPN and IG-semiTPN and IG-semiTPN are 59.34% vs. 53.13% on miniImageNet, and 64.80% vs. 56.09% on tieredImageNet, respectively. The results demonstrate the superior capacity of the proposed ensemble strategy in using the extra unlabeled information for boosting few-shot methods. Moreover, the addition of the distractor class enhances the robustness of the model.



Figure 12. Comparison of the IG-semiTPN and semi-HoETPN on miniImageNet.



Figure 13. Comparison of the IG-semiTPN and semi-HoETPN on tieredImageNet.

# 6. Conclusions and Future Work

Few-shot learning aims to construct a classification model using limited samples. In this paper, we propose a novel ensemble semi-supervised few-shot learning with a transductive propagation network and evidence fusion. During the process of transductive propagation learning, we introduce the preset weight coefficient and calculate the process of iterative inferences to present homogeneous and heterogeneous models to improve the stability of the model. Then, we propose the improved D-S evidence ensemble strategy to enhance the stability of the final results. It combines the information entropy to realize the pre-processing of the evidence source. Then, an improved ensemble pruning method adopting the L2 norm is proposed to maintain a better performance of individual learners to enhance the accuracy of model fusion. Furthermore, an interference set is introduced to improve the robustness of the semi-supervised model. Experiments on miniImagnet and tieredImageNet indicate that the proposed approaches outperform the state-of-the-art few-shot model. However, our proposal directly utilizes a label propagation approach to transfer information between nodes in the graph-constructing phase. Therefore, in our future work, we will consider adopting the reality-semantic and cross-modal information to improve the accuracy of the transduction inference graph in few-shot learning.

**Author Contributions:** Writing—original draft preparation, Conceptualization, Methodology, Software, validation, investigation, X.P.; writing—review and editing, supervision, funding acquisition, G.L.; writing—review and editing, Conceptualization, Formal analysis, funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is partly supported by the Nature Science Foundation of China under Grant (Nos. 60473125), Science Foundation of China University of Petroleum-Beijing At Karamay under Grant (Nos. RCYJ2016B-03-001), Kalamay Science & Technology Research Project (Nos. 2020CGZH0009), Natural Science Foundation of Fujian Province, China under Grant (Nos. 2021J011004 and 2021J011002), the Ministry of Education Industry-University-Research Innovation Program (Grant No. 2021LDA09003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

**Data Availability Statement:** The miniImageNet dataset and tieredImageNet dataset can be found at: https://github.com/renmengye/few-shot-ssl-public (accessed on 29 January 2023).

Acknowledgments: We are greatly indebted to colleagues at Data and Knowledge Engineering Center, School of Information Technology and Electrical Engineering, the University of Queensland, Australia. We thank Xiaofang Zhou, Xue Li, Shuo Shang and Kai Zheng for their special suggestions and many interesting discussions.

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

## References

- 1. Aouani, H.; Ayed, Y.B. Speech Emotion Recognition with deep learning. Procedia Comput. Sci. 2020, 176, 251–260. [CrossRef]
- 2. LeCun, Y.; Yoshua, B.; Hinton, G. Deep learning. Nature 2015, 521, 436–444. [CrossRef] [PubMed]
- Wang, C.; Wang, C.; Li, W.; Wang, H. A brief survey on RGB-D semantic segmentation using deep learning. *Displays* 2021, 70, 102080. [CrossRef]
- 4. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In Proceedings of the British Machine Vision Conference 2014, Nottingham, UK, 1–5 September 2014.
- 5. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 818–833.
- Chen, Z.; Fu, Y.; Zhang, Y.; Jiang, Y.G.; Xue, X.; Sigal, L. Semantic feature augmentation in few-shot learning. In Proceedings of the 5th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018.
- 7. Lu, J.; Li, J.; Yan, Z.; Mei, F.; Zhang, C. Attribute-based synthetic network (abs-net): Learning more from pseudo feature representations. *Pattern Recognit.* 2018, *80*, 129–142. [CrossRef]
- Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D. Matching networks for one shot learning. Adv. Neural Inf. Process. Syst. 2016, 29, 3630–3638.
- 9. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for fewshot learning. Adv. Neural Inf. Process. Syst. 2017, 30, 4077–4087.

- Xing, C.; Rostamzadeh, N.; Oreshkin, B.; Pinheiro, P.O. Adaptive cross-modal few-shot learning. *Adv. Neural Inf. Process. Syst.* 2019, 32, 4848–4858.
- 11. Lv, F.; Zhang, J.; Yang, G.; Feng, L.; Yu, Y.; Duan, L. Learning cross-domain semantic-visual relationships for transductive zero-shot learning. *Pattern Recognit.* 2023, 141, 109591. [CrossRef]
- Zhang, J.; Yang, G.; Hu, P.; Lin, G.; Lv, F. Semantic Consistent Embedding for Domain Adaptive Zero-Shot Learning, *IEEE Trans. Image Process.* 2023, 32, 4024–4035. [CrossRef]
- 13. Oreshkin, B.; López, P.R.; Lacoste, A. Tadam: Task dependent adaptive metric for improved few-shot learning. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 721–731.
- Tang, K.D.; Tappen, M.F.; Sukthankar, R.; Lampert, C.H. Optimizing one-shot recognition with micro-set learning. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2018; pp. 3027–3034.
- 15. Zheng, Y.; Wang, R.; Yang, J.; Xue, L.; Hu, M. Principal characteristic networks for few-shot learning. *J. Vis. Commun. Image Represent.* 2019, 59, 563–573. [CrossRef]
- Wang, D.; Zhang, M.; Xu, Y.; Lu, W.; Yang, J.; Zhang, T. Metric-based meta-learning model for few-shot fault diagnosis under multiple limited data conditions, *Mech. Syst. Signal Process.* 2021, 155, 107510. [CrossRef]
- 17. Rusu, A.A.; Rao, D.; Sygnowski, J.; Vinyals, O.; Pascanu, R.; Osindero, S.; Hadsell, R. Meta-learning with latent embedding optimization. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
- Jiang, X.; Havaei, M.; Varno, F.; Chartr, G.; Chapados, N.; Matwin, S. Learning to learn with conditional class dependencies. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
- Gidaris, S.; Komodakis, N. Generating classification weights with gnn denoising autoencoders for few-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 21–30.
- Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; Turner, R.E. Meta-learning probabilistic inference for prediction. In Proceedings
  of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
- 21. Bertinetto, L.; Henriques, J.F.; Torr, P.H.; Vedaldi, A. Meta learning with differentiable closed-form solvers. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
- 22. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
- 23. Jamal, M.A.; Qi, G.J. Task agnostic meta-learning for few shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long, Beach, CA, USA, 16–20 June 2019; pp. 11719–11727.
- Antoniou, A.; Edwards, H.; Storkey, A. How to train your maml. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
- Liu, Y.; Lee, J.; Park, M.; Kim, S.; Yang, E.; Hwang, S.J.; Yang, Y. Learning to propagate labels: Transductive propagation network for few-shot learning. In Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
- Huang, H.; Zhang, J.; Zhang, J.; Wu, Q.; Xu, C. PTN: A Poisson Transfer Network for Semi-supervised Few-shot Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 1602–1609.
- Rodríguez, P.; Laradji, I.H.; Drouin, A.; Lacoste, A. Embedding Propagation: Smoother Manifold for Few-Shot Classification. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 121–138.
- Iscen, A.; Tolias, G.; Avrithis, Y.; Chum, O. Label propagation for deep semisupervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5070–5079.
- Liu, B.; Wu, Z.; Hu, H.; Lin, S. Deep Metric Transfer for Label Propagation with Limited Annotated Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27October–2 November 2019; pp. 1317–1326.
- 30. Zhang, R.; Yang, S.; Zhang, Q.; Xu, L.; He, Y.; Zhang, F. Graph-based few-shot learning with transformed feature propagation and optimal class allocation. *Neurocomputing* **2022**, 470, 247–256. [CrossRef]
- Ma, Y.; Bai, S.; An S.; Liu, W.; Liu, A.; Zhen, X.; Liu, X. Transductive Relation-Propagation Network for Few-shot Learning. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20), Yokohama, Japan, 7–15 January 2021; pp. 804–810.
- 32. Simon, C.; Koniusz, P.; Nock, R.; Harandi, M. Adaptive Subspaces for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Seattle, WA, USA, 13–19 June 2020; pp. 4135–4144.
- Kim, J.; Kim, T.; Kim, S.; Yoo, C.D. Edge-Labeling Graph Neural Network for Few-Shot Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11–20.
- Li, X.; Huang, J.; Liu, Y.; Zhou, Q.; Zheng, S.; Schiele, B.; Sun, Q. Learning to teach and learn for semi-supervised few-shot image classification. *Comput. Vis. Image Underst.* 2021, 212, 103270. [CrossRef]
- 35. Ayyad, A.; Li, Y.; Muaz, R.; Albarqouni, S.; Elhoseiny, M. Semi-Supervised Few-Shot Learning with Prototypical Random Walks; arXiv 2019, arXiv:1903.02164.
- 36. Huang, S.; Zeng, X.; Wu, S.; Yu, Z.; Azzam, M.; Wong, H.S. Behavior regularized prototypical networks for semi-supervised few-shot image classification. *Pattern Recognit.* **2021**, *112*, 107765. [CrossRef]

- Dvornik, N.; Mairal, J.; Schmid, C. Diversity With Cooperation: Ensemble Methods for Few-Shot Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3722–3730.
- Zhou, M.; Li, Y.; Lu, H. Ensemble-Based Deep Metric Learning for Few-Shot Learning. In Proceedings of the 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, 15–18 September 2020; pp. 406–418.
- 39. Yu, T.; He, S.; Song, Y.Z.; Xiang, T. Hybrid Graph Neural Networks for Few-Shot Learning. In Proceedings of the AAAI—Thirty-Eighth Conference on Artificial Intelligence, Vancouver, BC, USA, 20–27 February 2022; pp. 3179–3187.
- 40. Liu, Y.; Schiele, B.; Sun, Q. An Ensemble of Epoch-Wise Empirical Bayes for Few-Shot Learning. In Proceedings of the 16th European Conference, Glasgow, UK, 23–28 August 2020; Volume 16, pp. 404–421.
- Pan, X.; Li, G.; Yu, Q.; Guo, K.; Li, Z. Novel Graph Semi-Supervised Transduction Approach with Improved Gauss Kernel for Few-Shot Learning. *Comput. Eng. Appl.* 2023, 59, 328–333.
- Ren, M.; Triantafillou, E.; Ravi, S.; Snell, J.; Swersky, K.; Tenenbaum, J.B.; Larochelle, H.; Zemel, R.S. Meta-learning for semisupervised few-shot classification. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
- Yu, Z.; Chen, L.; Cheng, Z.; Luo, J. TransMatch: A Transfer-Learning Scheme for Semi-Supervised Few-Shot Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12856–12864.
- Greenleaf, G.; Mowbray, A.; King, G.; Cant, S.; Chung, P. More than wyshful Thinking: AustLII's Legal Inferencing via the World Wide Web. In Proceedings of the ICAIL97: International Conference on Artificial Intelligence and Law, Melbourne, Australia, 30 June–3 July 1997; pp. 47–55.
- 45. Krogh, A.; Vedelsby, J. Neural Network Ensembles, Cross Validation, and Active Learning. In Proceedings of the International Conference on Neural Information Processing Systems, Denver, CO, USA, 27 November–2 December 1995.
- 46. Li, N.; Zhou, Z.H. Selective Ensemble under Regularization Framework; Springer: Berlin/Heidelberg, Germany, 2009.
- 47. Dempster, A.P. Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **1967**, *38*, 325–339. [CrossRef]
- 48. Shafer, G. A Mathematical Theory of Evidence; Princeton University Press: Princeton, NJ, USA, 1976.
- 49. Xiao, J.; Tong, M.; Zhu, C.; Fan, Q. Improved combination rule of evidence based on pignistic probability distance. *J. Shanghai Jiaotong Univ.* **2012**, *46*, 636–641+645.
- 50. Deng, Y.; Shi, W.; Zhu, Z. Efficient combination approach of conflict evidence, in Chinese. J. Infr. Millim. Waves 2004, 23, 27–32.
- 51. Choi, E.; Lee, C. Feature extraction based on the Bhattacharyya distance. Pattern Recognit. 2003, 36, 1703–1709. [CrossRef]
- Satorras, V.G.; Estrach, J.B. Few-Shot Learning with Graph Neural Networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
- Luo, ; Y.; Huang, Z.; Zhang, Z.; Wang, Z.; Baktashmotlagh, M.; Yang, Y. Learning from the Past: Continual Meta-Learning via Bayesian Graph Modeling. In Proceedings of the AAAI—Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020.
- Yang, L.; Liangliang, L.; Zilun, Z.; Xinyu, Z.; Erjin, Z.; Yu, L. DPGN: Distribution Propagation Graph Network for Few-Shot Learning. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13387–13396.
- 55. Chen, H.; Li, H.; Li, Y.; Chen, C. Sparse spatial transformers for few-shot learning. Sci. China Inf. Sci. 2023, 66, 210102. [CrossRef]
- 56. Jia, X.; Su, Y.; Zhao, H. Few-shot learning via relation network based on coarse-grained granulation. *Appl. Intell.* **2023**, *53*, 996–1008. [CrossRef]
- 57. Su, Y.; Zhao, H.; Lin, Y. Few-shot learning based on hierarchical classification via multi-granularity relation networks. *Int. J. Approx. Reason.* **2022**, *142*, 417–429. [CrossRef]
- 58. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
- Kingma, D.P.; Jimmy, B. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
- 60. Chang, C.C.; Li, C.J. LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2011, 2, 1–39. [CrossRef]
- 61. Hsu, C.; Chang, C.; Lin, C. A practical guide to support vector classification, BJU Int. 2008, 101, 1396–1400.
- Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P.H.; Hospedales, T.M. Learning to compare: Relation network for few-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
- Qiao, S.; Liu, C.; Shen, W.; Yuille, A.L. Few-Shot Image Recognition by Predicting Parameters from Activations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 7229–7238.

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