electronics

MDPI

Article

Discrepant Semantic Diffusion Boosts Transfer Learning Robustness

Yajun Gao [1,†], Shihao Bai [2,†], Xiaowei Zhao [1], Ruihao Gong [1,2], Yan Wu [3] and Yuqing Ma [1,4,*]

[1] State Key Lab of Software Development Environment, Beihang University, Beijing 100191, China; yajungao@buaa.edu.cn (Y.G.); xiaoweizhao@buaa.edu.cn (X.Z.); gongruihao@sensetime.com (R.G.)
[2] SenseTime Research, Beijing 100080, China; baishihao@sensetime.com
[3] Beijing Academy of Science and Technology, Beijing 100089, China; wuy@bjast.ac.cn
[4] Institute of Artifical Intelligence, Beihang University, Beijing 100191, China
* Correspondence: mayuqing@buaa.edu.cn
† These authors contributed equally to this work.

Abstract: Transfer learning could improve the robustness and generalization of the model, reducing potential privacy and security risks. It operates by fine-tuning a pre-trained model on downstream datasets. This process not only enhances the model's capacity to acquire generalizable features but also ensures an effective alignment between upstream and downstream knowledge domains. Transfer learning can effectively speed up the model convergence when adapting to novel tasks, thereby leading to the efficient conservation of both data and computational resources. However, existing methods often neglect the discrepant downstream–upstream connections. Instead, they rigidly preserve the upstream information without an adequate regularization of the downstream semantic discrepancy. Consequently, this results in weak generalization, issues with collapsed classification, and an overall inferior performance. The main reason lies in the collapsed downstream–upstream connection due to the mismatched semantic granularity. Therefore, we propose a discrepant semantic diffusion method for transfer learning, which could adjust the mismatched semantic granularity and alleviate the collapsed classification problem to improve the transfer learning performance. Specifically, the proposed framework consists of a Prior-Guided Diffusion for pre-training and a discrepant diffusion for fine-tuning. Firstly, the Prior-Guided Diffusion aims to empower the pre-trained model with the semantic-diffusion ability. This is achieved through a semantic prior, which consequently provides a more robust pre-trained model for downstream classification. Secondly, the discrepant diffusion focuses on encouraging semantic diffusion. Its design intends to avoid the unwanted semantic centralization, which often causes the collapsed classification. Furthermore, it is constrained by the semantic discrepancy, serving to elevate the downstream discrimination capabilities. Extensive experiments on eight prevalent downstream classification datasets confirm that our method can outperform a number of state-of-the-art approaches, especially for fine-grained datasets or datasets dissimilar to upstream data (e.g., 3.75% improvement for Cars dataset and 1.79% improvement for SUN dataset under the few-shot setting with 15% data). Furthermore, the experiments of data sparsity caused by privacy protection successfully validate our proposed method's effectiveness in the field of artificial intelligence security.

Keywords: transfer learning; semantic diffusion; model robustness

Citation: Gao, Y.; Bai, S.; Zhao, X.; Gong, R.; Wu, Y.; Ma, Y. Discrepant Semantic Diffusion Boosts Transfer Learning Robustness. *Electronics* **2023**, *12*, 5027. https://doi.org/10.3390/electronics12245027

Academic Editor: Fabio Grandi

Received: 15 November 2023
Revised: 14 December 2023
Accepted: 14 December 2023
Published: 16 December 2023

1. Introduction

Transfer learning could mitigate the potential privacy and security risks through effectively re-utilizing the pre-trained model, which is initially trained on a voluminous and information-dense dataset, commonly referred to as the upstream dataset. The pre-trained model is adaptively fine-tuned on a novel and less-data-intensive dataset, denominated as the downstream dataset. For example, in special critical scenarios where data are

seldom publicly available, transfer learning can reduce the training data requirement for models on new tasks, thereby lowering the potential security risks [1–4]. It is capable of accelerating the learning of downstream models, while reducing the labor costs and computational consumption. Moreover, pre-trained models, which are typically public and thus more vulnerable to attacks, have been shown to benefit from the security enhancements provided by transfer learning. For example, studies [5–7] examining the susceptibility of downstream models to attacks have confirmed that transfer learning can protect these downstream models from being easily attacked, thereby enhancing their robustness. Due to its practicality, transfer learning has attracted extensive attention in the field of computer vision [8–12], and has been applied in many task scenarios such as transportation, medical treatment [13], social media [14] and art [15–18]. Various works [19] have been proposed to explore the problem of what and how to transfer from the pre-trained models.

Transfer learning allows the use of a pre-trained model's comprehensive upstream knowledge to better accomplish the downstream task, even with limited labeled data. By aligning the knowledge domain, model convergence in downstream tasks is accelerated. Some attempts have been made to explore and apply the shared downstream–upstream knowledge alignment through the well-designed regularization in the fine-tuning stage. Specifically, Refs. [20,21] tried to add the regularization of parameters or features in the fine-tuning stage to preserve the diversity of pre-training information, and thus to selectively transfer the upstream knowledge of the pre-trained network. Ref. [22] adaptively selected layers to freeze or fine-tune according to each training sample. To preserve as much upstream information as possible, Ref. [23] placed constraints on the weights of the fine-tuned model to approximate those of the pre-trained model. Similarly, Ref. [24] supervised the fine-tuning process with both upstream probabilistic labels and downstream labels. Meanwhile, there are also a few works [25–27] aiming to boost the generalization ability of pre-trained feature representations.

Although previous transferring methods have achieved satisfactory performance, they neglected the discrepant downstream–upstream connections and rigidly preserved the upstream information without an adequate regularization of downstream semantic discrepancy, leading to weak generalization, the collapsed classification problem, and inferior transferring performance. Discrepant downstream–upstream connections pertain to the mapping of semantics from downstream classes to upstream classes, aiming to preserve the semantic diversity and discrimination in downstream data. As shown in Figure 1, the connection marked with a check symbol after 'discrepant' is key to maintaining feature distinctiveness in the downstream embedding space, a process we also describe as ensuring downstream semantic discrepancy. As shown in Figure 1a, previous methods assumed that the downstream classes and upstream classes share similar semantic granularity (a one–one downstream–upstream mapping), where the downstream–upstream connection of each downstream category is discrepant from the others, and thus concluded that simply preserving the upstream information will benefit the downstream discrimination. However, for the fine-grained dataset or the dataset of low similarity with the upstream dataset as shown in Figure 1b, if a downstream category is only allowed to connect a single upstream class, it is very likely that an upstream class simultaneously contains the representative characteristics of different downstream categories, due to the mismatched downstream–upstream granularity (a many–one downstream–upstream mapping). In this condition, the fine-tuned model, which is rigidly restricted to preserve the upstream information, forces the representations of those downstream classes close to each other and thus counters to the downstream discriminative classification task, resulting in performance degradation.

To address the above problems as shown in the right column in Figure 1c, we should encourage that a downstream class diffusively relates to k ($k > 1$) upstream classes rather than a single one during transferring (for example, $k = 2$ in the figure), introducing more semantic diversity and discrimination to adjust the mismatched downstream–upstream connection, meanwhile ensuring the discrepancy of these downstream–upstream connections for different downstream classes to achieve robust classification. Therefore, we

propose a discrepant semantic diffusion method for transfer learning, which could ensure the downstream semantic discrepancy to effectively accomplish downstream tasks. This method encourages semantic diffusion, a process where knowledge from the upstream task is diffused in a controlled and differentiated manner to the downstream task to avoid collapsed classification.

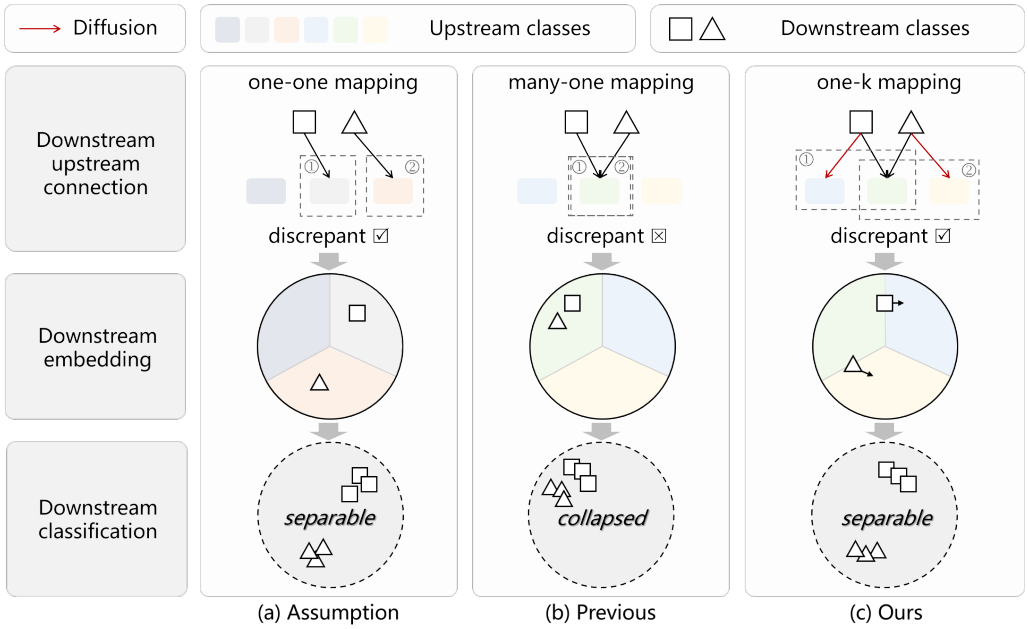

Figure 1. Illustration of the discrepant semantic diffusion. (**a**) Assumption: downstream classes and upstream classes share similar semantic granularity, and their connections could be represented as a one–one downstream–upstream mapping, which ensures downstream semantic discrepancy and makes downstream classification separable. (**b**) Previous: due to many–one downstream–upstream mapping, features of different downstream categories are embedded in the same upstream space, leading to collapsed classification problem. (**c**) Ours: the discrepant semantic diffusion method is proposed to realize discrepant one–*k* downstream–upstream mapping to ensure downstream semantic discrepancy, which avoids the collapsed classification problem and achieves robust separation of the downstream classification.

Specifically, the proposed discrepant semantic diffusion contains Prior-Guided Diffusion (PGD) in the pre-training stage and Discrepant Diffusion (DI2) in the fine-tuning stage. PGD relaxes the centralized one-hot classification regularization during the pre-training stages through a semantic prior, which is a predefined knowledge structure that enables the model to understand potential semantic associations in the data. PGD could encourage the following diffusive one-*k* downstream–upstream mapping during the transferring and improve the generalization ability of the pre-trained model. The semantic prior indicates the semantic similarity between two classes, according to which we could reduce the penalty for classes similar to the ground-truth class during the gradient back-propagation, and thus sets the foundation for top-*k* mapping during the fine-tuning stage. Moreover, DI2 adjusts the semantic granularity by encouraging diffusive downstream–upstream connection during the fine-tuning stage, and regularizes the diffusive connection via the downstream semantic discrepancy, effectively and adaptively transferring the upstream information to accomplish the downstream task. This approach introduces more semantic diversity with the one–*k* downstream–upstream mapping, maintaining a different granularity diffusion sequence for each downstream class, rather than collapsing into the same category. By discreetly determining the value of *k*, downstream discrimination is ensured from the downstream–upstream connection perspective. It is noteworthy that for downstream datasets with varying semantic granularities, the value of *k* differs, endowing the model

with a continuous adaptability capability. Consequently, DI2 could guarantee effective transferring specific to the downstream task.

To summarize, we build the diffusive downstream–upstream connection with the semantic discrepancy in transfer learning. With the discrepant semantic diffusion, the proposed model could effectively avoid the collapsed classification problem to boost transfer learning robustness and improve performance on downstream tasks. The comprehensive experiments on eight downstream classification datasets demonstrate the superiority and generalization of our method compared to the state-of-the-art transfer learning models, especially for the fine-grained datasets or datasets, unlike upstream data (e.g., 3.75% improvement for the Cars dataset and 1.79% improvement for the SUN dataset). Furthermore, a key aspect of our method is its impact on privacy and security risks, especially during the fine-tuning process on downstream datasets. As the model adapts to new tasks, there is a potential risk of exposing sensitive information embedded in the training data [28,29]. This risk is particularly significant in scenarios involving downstream datasets that contain private or confidential information. Therefore, it is necessary to study the performance of our method in scenarios of data sparsity caused by privacy protection. We explore the data sparsity issue caused by privacy protection in the field of artificial intelligence (AI) security, successfully validating our proposed method's effectiveness through experiments. We make the following contributions:

- We propose a transfer learning method with discrepant semantic diffusion to better adapt to various downstream datasets, diffusing more upstream in-depth semantic discrepancy information into the downstream dataset categories.
- To empower the pre-trained model with the semantic diffusion ability, Prior-Guided Diffusion (PGD) is introduced into the pre-training stage, which relaxes the centralized one-hot classification regularization through the semantic prior.
- To further promote the downstream discrimination, Discrepant Diffusion (DI2) is designed to maintain a different granularity diffusion sequence for each downstream category, which guides the model to focus on discriminative information among the downstream categories.
- Experimental results on eight prevalent downstream classification datasets and various networks verify the effectiveness of the proposed method. The additional experiments demonstrate that our method significantly reduces the training data requirements for models on new tasks and lowers potential security risks.

2. Related Work

ImageNet pre-trained deep neural networks have shown remarkable transferability to various tasks, such as image classification [30–32], image segmentation [33–36], object detection [37–40], image retrieval [41], action recognition [42,43], etc. Even in cases where upstream and downstream tasks have significant differences, such as transfer for depth estimation [44], medical imaging [45–47] and other downstream tasks, ImageNet pre-trained deep neural networks are also effective and could speed up the generalization proved by [25].

Following the typology of [48], from the perspective of transferring and aligning robust knowledge, the current transfer learning methods are mainly twofold, knowledge of feature representations transfer and knowledge of parameters (i.e., inductive bias) transfer. For feature representation transfer learning, researchers aim to improve the generalization and transferability of pre-trained feature representations. Through a thorough study of transfer learning, Refs. [25,26] pointed out it is mainly low-level and mid-level general representations that are transferred instead of specialized high-level representations. Follow the above findings, Ref. [27] increases the transferability of high-level feature using adversarial robust pre-trained networks, while [49] obtains representations with higher generality by pre-training on large-scale upstream datasets. Some methods use contrast learning or self-supervised learning during model pre-training to achieve better generalization ability. Using nearest-neighbor contrastive learning, simple-architecture contrastive learning and

cross-view alignment contrastive learning, respectively, Refs. [50–52] enhances the visual representations to improve the transfer learning performance. Ref. [53] enables the control of upstream performance and transferability through a self-supervised learning setup.

Inductive transfer tends to regularize the weights of the fine-tuning model with the pre-trained weights, which allows the model to leverage previously learned knowledge, aligning knowledge domains to enhance robustness. Since the parameters preserve the inductive bias gained from the upstream tasks, fine-tuned parameter values ending up far from the pre-trained ones are regarded as a sign of forgetting [54]. Therefore, inductive transfer methods mostly focus on retaining the information mined from upstream datasets throughout the fine-tune process. For example, to transfer more upstream information, Ref. [23] regularizes the weights of the fine-tuned model with L2 constraints to make it approximate to the weights of the pre-trained model. To selectively preserve the upstream information, Ref. [20] penalizes deviations of network activation. For similar purposes, Ref. [21] penalizes small eigenvalues in representations that cause negative transfer.

Though ImageNet is still the de facto pre-train dataset [55], not all knowledge gained from its domain and task is as informative and universal for downstream tasks. Forcing the model to simply retain as much knowledge as possible may sometimes degrade the performance [56]. On the representation side, the fixed-feature setting where a linear classifier is trained on top of the pre-trained extractor is usually outperformed by fine-tuning the full model. Even when the robustness of pre-trained representations is improved, Ref. [27] suggests that ImageNet pre-trained networks still work better as the weight initialization instead of feature extractors. Ref. [55] found out that higher accuracy on ImageNet does not guarantee better transfer performance, which indicates there might be some disagreement between the discrimination among upstream and downstream categories. And on the inductive side, Ref. [56] revealed that weight preservation would sometimes hinder the transfer, especially when downstream domains diverge from ImageNet. Besides the domain differences, analysis from [26] also shows that the task misalignment between ImageNet and downstream datasets weakens the transferability. Therefore, to improve transfer learning robustness and better adapt to various downstream datasets, it is crucial to distinguish and align the useful part of the knowledge and discard the disadvantaging ones.

3. The Approach

Previous transfer learning approaches neglect the valuable semantic discrepancy of the downstream–upstream connection during the transferring process, leading to weak generalization and collapsed classification, especially on downstream datasets, unlike the upstream dataset, as we mentioned in Section 1. To handle this problem, in this paper, we propose a robust transfer learning method with discrepant semantic diffusion, which could introduce the semantic discrepancy during the transferring process through the semantic diffusion, building the discrepant one–many downstream–upstream connection for different downstream classes. Therefore, we first propose Prior-Guided Diffusion (PGD) for pre-training to encourage the semantic diffusion, and then introduce Discrepant Diffusion (DI2) for fine-tuning to regularize the transferring with a diffusive downstream–upstream connection constrained by the semantic discrepancy. In this way, we could generalize and transfer the effective upstream information, especially for downstream tasks, improving model generalization and performance. Figure 2 illustrates the architecture of the proposed method.

In the following sections, we will first introduce the preliminary and discuss the possible reason for the inferior performance of previous methods. Then, we will elaborate the details of the proposed PGD for pre-training and DI2 for fine-tuning. Finally, we will present the whole learning pipeline.

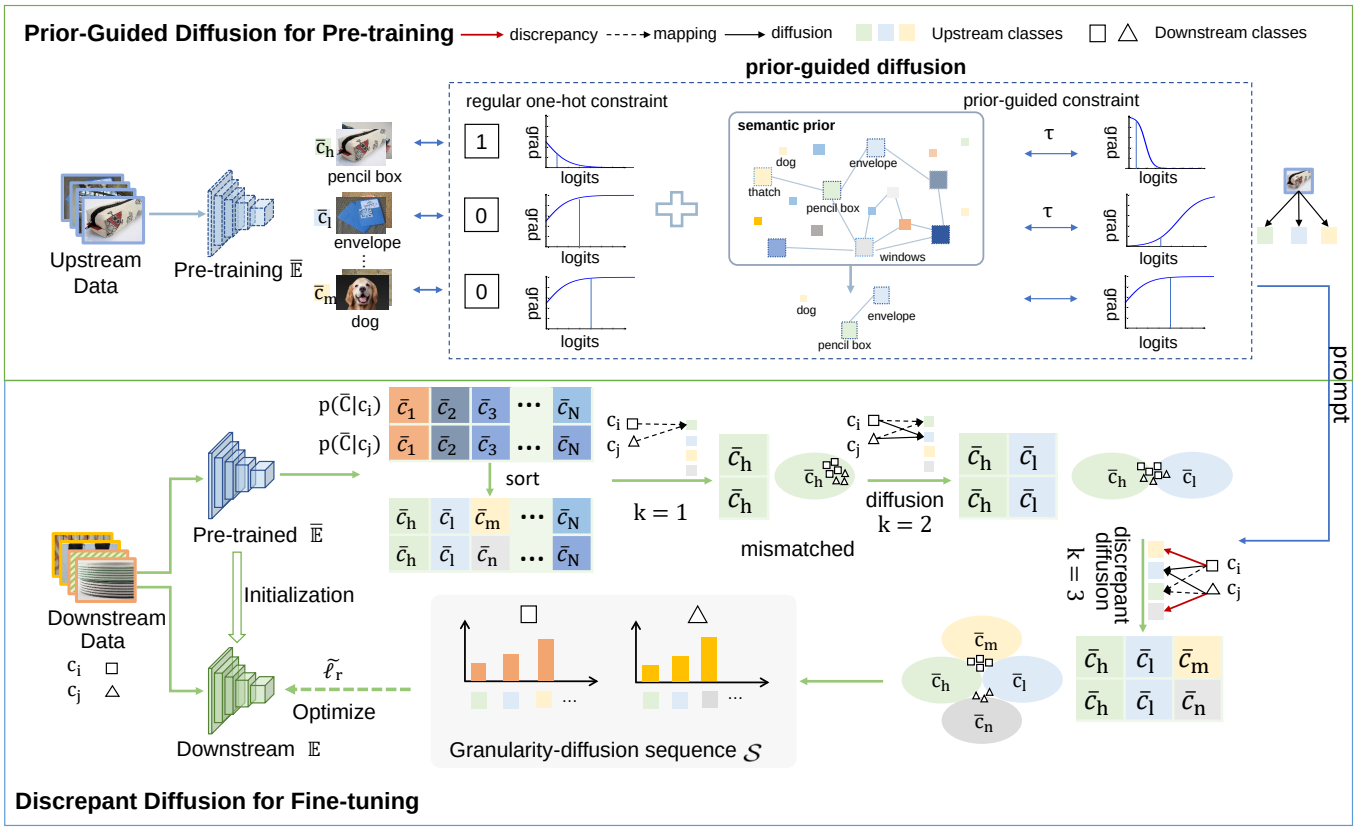

Figure 2. The framework of our discrepant semantic diffusion method for transfer learning consists of Prior-Guided Diffusion (PGD) for pre-training and Discrepant Diffusion (DI2) for fine-tuning. Our method first conducts the pre-training with the semantic prior, enabling the semantic diffusion for further fine-tuning. Then the discrepant diffusion in the fine-tuning stage constructs a discrepant and diffusive downstream–upstream connection $\mathcal{S}$ based on the sorted averaged logits of the pretrained model $p(\bar{\mathbf{C}}|c_i)$,$p(\bar{\mathbf{C}}|c_j)$ of the i-th and j-th downstream classes, effectively alleviating the collapsed classification problem and improving the downstream discrimination.

3.1. Preliminary

In this section, we will give a formal description of the transfer learning pipeline. The standard transfer learning process usually contains a pre-training stage, where a network is trained on a large-scale upstream dataset and a fine-tuning stage, where the pre-trained network will be fine-tuned adapted to the downstream tasks. We first denote $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}$ as the downstream dataset, where $\mathbf{x}_i$ is the i-th data sample, while $\mathbf{y}_i \in \mathcal{C}$ is its corresponding class label. $\mathcal{C}$ is the class set containing M classes $\mathcal{C} = \{c_1, c_2, ..., c_M\}$. Similarly, $\bar{\mathcal{D}} = \{\bar{\mathbf{x}}_i, \bar{\mathbf{y}}_i\}$ is the upstream dataset, where $\bar{\mathbf{x}}_i$ is the i-th data sample, while $\bar{\mathbf{y}}_i \in \bar{\mathcal{C}}$ is its corresponding class label. $\bar{\mathcal{C}}$ is the class set containing N classes $\bar{\mathcal{C}} = \{\bar{c}_1, \bar{c}_2, ..., \bar{c}_N\}$. Normally, $N \gg M$, meaning that the upstream dataset is more informative and the model trained on such a dataset is empowered with strong generalizable representation learning capability. Therefore, we would like to take advantage of a model with strong robustness and generalization and apply it to the downstream task, avoiding the time-consuming training on the downstream dataset from scratch.

To transfer the strong representation ability of the model $\bar{\mathbb{G}} = \{\bar{\mathbb{E}}, \bar{\mathbb{FC}}\}$ pre-trained on $\bar{\mathcal{D}}$, where $\bar{\mathbb{E}}$ is the feature extractor and $\bar{\mathbb{FC}}$ is the task-specific classifier, the most intuitive practice is replacing the original classifier with the new classifier adapted to the downstream task and fine-tuning the whole network $\mathbb{G} = \{\mathbb{E}, \mathbb{FC}\}$:

$$\min \frac{1}{|\mathcal{D}|} \sum_{i=0}^{|\mathcal{D}|} \ell_c(\sigma(\mathbb{G}(\mathbf{x}_i)), \mathbf{y}_i) + \lambda \ell_r(\cdot), \tag{1}$$

where ℓ_c is the multi-classification loss and σ is the sigmoid activation function. λ is the weighting hyper-parameter. $\ell_r(\cdot)$ is usually the regularization term forcing the fine-tuned network to preserve the appropriate upstream information. For instance, Ref. [24] models the downstream–upstream matching $p(\bar{C}|c_m)$ through averaging the predictions of the pre-trained model over all samples of each downstream category c_m, and thus encourages the output of fine-tuned model $\mathbb{E}$ for each sample of c_m closed to the $p(\bar{C}|c_m)$:

$$\ell_r = \sum_{m} \sum_{\{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D} | \mathbf{y}_i = m\}} \ell_c(\bar{\mathbb{F}}\bar{\mathbb{C}}(\mathbb{E}(\mathbf{x}_i)), p(\bar{C}|c_m)). \tag{2}$$

Although these methods transfer upstream information rapidly and achieve satisfactory performance on the downstream task, without sufficiently considering the discrepant downstream–upstream connection during transferring on different downstream dataset, it leads to insufficient robustness and poor generalization of the model.

3.2. Prior-Guided Diffusion for Pre-Training

The pre-training model of traditional transfer learning methods is usually supervised by the centralized one-hot labels, posing a many–one downstream–upstream connection risk for collapsed classification when generalizing and transferring to other datasets of dissimilar distribution. The semantic diffusion could encourage a sample to relate to different classes rather than its ground-truth class, and thus will introduce more semantic diversity when transferring. Therefore, we present a Prior-Guided Diffusion approach for pre-training, introducing a semantic prior which indicates the precious class relations and potential semantic associations, thereby guiding diffusive one–k downstream–upstream mapping. According to the precious semantic prior, we could select the neighboring classes of the ground-truth class and reduce the penalty for the misclassification on them. Moreover, we also maintain the discrimination ability of the pre-trained model, assigning the largest prediction score for the ground-truth class. The PGD could enhance the generalization and diffusion ability of the pre-trained model, exploring common semantic patterns rather simply concentrating on discriminative ones, and thus could benefit the transferring and promote the downstream semantic discrepancy in the following fine-tuning stage.

The semantic prior, such as additional teacher work (such as the prevalent transformer [57,58]), or an artificially defined WordNet synonym structure tree, is introduced to better model the category connection. It reflects the similarity degree between different classes from the semantic perspective. For an upstream sample $\bar{\mathbf{x}}_i$, we denote the function $\mathcal{N}(\bar{\mathbf{y}}_i)$ to return the similar upstream category sequences for the ground-truth upstream category $\bar{\mathbf{y}}_i$ according to the prior. Specifically, we introduce a hyper-parameter τ to the n-th element of the logits $\mathbf{l}_i = \mathbb{G}(\mathbf{x}_i)$, diminish the penalty for this class, and thus encourage a similar connection:

$$\begin{aligned} \ell_u = -\sum_{i=0}^{|\bar{\mathcal{D}}|} \sum_{n=0}^{N-1} (\bar{\mathbf{y}}_i[n] \log(\sigma(\mathbf{l}_i[n] - \mathcal{I}(\bar{c}_n \in \mathcal{N}(\bar{\mathbf{y}}_i)) \log \tau)) \\ + (1 - \bar{\mathbf{y}}_i[n]) \log(1 - \sigma(\mathbf{l}_i[n] - \mathcal{I}(\bar{c}_n \in \mathcal{N}(\bar{\mathbf{y}}_i)) \log \tau))) \end{aligned} \tag{3}$$

where $\mathcal{I}(\cdot)$ is the Kronecker delta function that is equal to 1 when the input condition holds, and 0 otherwise. We adopt Binary Cross Entropy (BCE) loss to guide the learning.

Theoretically speaking, the absolute value of the loss gradient could be divided into the following conditions:

$$\begin{cases} \dfrac{\tau}{\tau + e^{\mathbf{l}_i[n]}}, & \bar{\mathbf{y}}_i[n] = 1 \\ \dfrac{e^{\mathbf{l}_i[n]}}{\tau + e^{\mathbf{l}_i[n]}}, & \bar{\mathbf{y}}_i[n] = 0, c_n \in \mathcal{N}(\bar{\mathbf{y}}_i) \\ \dfrac{e^{\mathbf{l}_i[n]}}{1 + e^{\mathbf{l}_i[n]}}, & \bar{\mathbf{y}}_i[n] = 0, c_n \notin \mathcal{N}(\bar{\mathbf{y}}_i) \end{cases} \tag{4}$$

Since $\tau \gg 1$, this gradient increases the penalty of the ground-truth class, and reduces the network's suppression of the wrong but similar classes while maintaining the penalty of the others. Hence, it could explore the valuable category correlation through the semantic prior, encouraging the semantic diffusion and the robustness of the pre-trained model.

PGD could assist the semantic diffusion in DI2, providing an advanced initialization for DI2 optimization. Moreover, building valuable class relations could help the pre-trained model explore the general semantic attributes rather than the discriminative ones for classification, improving the generalization ability cost effectively. Consequently, simply adopting our PGD without the semantic discrepancy constraint in DI2 will also achieve a good performance.

3.3. Discrepant Diffusion for Fine-Tuning

We further introduce Discrepant Diffusion in the fine-tuning stage, and build the diffusive downstream–upstream connection with semantic discrepancy, which could diffusively and discriminatively transfer the appropriate upstream information and enhance the downstream discrimination. We will first illustrate the semantic diffusion between upstream and downstream classes, and then demonstrate how the downstream semantic discrepancy regularizes the diffusion.

Semantic diffusion: The semantic diffusion could effectively adjust the mismatched downstream–upstream granularity through a diffusive downstream–upstream connection, alleviating the collapsed classification problem. To explicitly model the downstream–upstream connection, we refer to the work [24] to calculate the average logits $p(\bar{C}|c_m) \in \mathcal{R}^{1\times N}$ on N upstream classes for each downstream category c_m. Instead of connecting a downstream class to the most similar upstream class to transfer the upstream information, we select top-k similar upstream categories for each downstream category according to $p(\bar{C}|c_m) \in \mathcal{R}^{1\times N}$, forming *granularity diffusion sequence* $\mathcal{S}^m$, where $|\mathcal{S}^m| = k$. Diffusing to multiple upstream classes could introduce more semantic diversity through the one–k mapping, and thus provide the possibility of discrepancy for downstream samples of different classes, rather than collapsing into the same category.

Semantic discrepancy: Although semantic diffusion could alleviate the collapsed classification problem, it cannot ensure downstream discrimination, which is the objective of transfer learning. Hence, we should further constrain the diffusion process towards the downstream discrimination. The semantic discrepancy is introduced to the diffusion through tuning the k-value. This adjustment, maintaining consistency with our earlier definition of k, ensures that the granularity diffusion sequence of an arbitrary downstream category is different from the counterparts of the others. In other words, for each downstream dataset, the k-value is adaptively determined according to the inherent discrepancy of the downstream dataset itself, which is totally consistent with our intuition. For example, the "banded" and "lined" categories in the DTD dataset are significantly different from the upstream ImageNet dataset, and even the top four adjacent upstream categories are the same, which makes it hard to ensure the discrimination. However, the fifth adjacent upstream categories are different, and $k = 5$ for other downstream category pairs also satisfies the condition. Therefore, $k = 5$ is appropriate for the DTD dataset. We list the different values of k for the downstream datasets in Figure 3, which also reflects the downstream–upstream semantic granularity differences.

As a result, the regularization term, which forces the fine-tuned network to preserve the appropriate upstream information and considers the discrepant semantic diffusion, could be written as:

$$\begin{aligned} \widetilde{\ell}_r = -\sum_{m=0}^{M} \sum_{\{(\mathbf{x}_i,\mathbf{y}_i)\in\mathcal{D}|\mathbf{y}_i=m\}} \sum_{n=0}^{N-1} (\mathcal{I}(c_n \in \mathcal{S}^m) \log \mathbf{h}_i^m[n] + \\ (1-\mathcal{I}(c_n \in \mathcal{S}^m)) \log(1-\mathbf{h}_i^m[n])) \\ s.t.\ \ (\mathcal{S}^u - \mathcal{S}^u \cap \mathcal{S}^v) \neq \phi,\ 0 \leq u, v \leq M, u \neq v \end{aligned} \tag{5}$$

where $\mathbf{h}_i^m = \sigma(\bar{\mathbb{F}}\bar{\mathbb{C}}(\mathbb{E}(\mathbf{x}_i)))$, and $\mathbf{h}_i^m[n]$ means the n-th elements of the prediction $\mathbf{h}_i^m$.

Figure 3. The k-value (namely the length of the granularity diffusion sequence $\mathcal{S}^m$ for the m-th downstream class) for different downstream datasets with the ImageNet as the upstream dataset. The k-value is dataset specific, reflecting the semantic granularity difference between an upstream dataset and the downstream dataset.

The semantic discrepancy could restrict the semantic diffusion towards a reasonable direction, selectively transferring the upstream information for different downstream categories of different downstream datasets, improving the transferring ability and the downstream performance.

Combined with the semantic diffusion and semantic discrepancy, DI2 could construct a diffusive downstream–upstream connection and thus adjust the mismatched semantic granularity. Meanwhile, it ensures the semantic discrepancy of the downstream tasks, further improving the transfer learning robustness. DI2 could effectively take advantage of the abundant upstream information and focus more on the discriminative characteristics of the downstream categories to avoid the collapsed classification. However, DI2 could also lead to a limitation in that our method is more suited for handling downstream tasks with fine-grained datasets, and tends to be less effective on coarser-grained downstream datasets.

3.4. Learning Pipeline

The proposed transfer learning method first conducts the pre-training with the semantic prior, preparing for the further diffusion in the fine-tuning stage. We determine the output of $\mathcal{N}$ for each upstream category through the prior, and thus utilize the prior to guide the pre-training through ℓ_u. After that, we obtain a more generalized pre-trained model empowered with the semantic diffusion ability.

Then, the discrepant diffusion for fine-tuning trains the model adaptive to the downstream classification tasks through the binary cross-entropy loss ℓ_d:

$$\begin{aligned}\ell_d = -\sum_{i}^{|\mathcal{D}|}\sum_{m}^{M}(\mathbf{y}_i[m]\log(\sigma(\mathbb{G}(\mathbf{x}_i)[m]))+ \\ (1-\mathbf{y}_i[m])\log(1-\sigma(\mathbb{G}(\mathbf{x}_i)[m]))),\end{aligned} \tag{6}$$

and regularizes the upstream information, preserving it as $\widetilde{\ell}_r$ in Equation (5). The overall loss for the fine-tuned model is:

$$\ell_o = \ell_d + \lambda\widetilde{\ell}_r, \tag{7}$$

where λ is the weighting hyper-parameter.

In the end, we could obtain the fine-tuned model adaptive to each downstream task, speeding up the convergence of the network with limited labeled data and enhancing the downstream performance. The computation of our algorithm is not complicated, so it can be implemented at a relatively low cost on the backbone network and does not require expensive computing resources.

4. Experiments

In this section, we evaluate the proposed model on widely used benchmarks compared with the state-of-the-art transfer learning approaches. We will first introduce the experimental setting. Then, we will illustrate the comparison results with the state of the art under the full-dataset and few-shot settings, respectively. After that, we will investigate each component of our model to verify their effectiveness. We will also examine the transfer performance on different model architectures and different pre-training methods. Finally, we will present rich experimental results for hyper-parameter analysis.

4.1. Experimental Setting

Dataset: The upstream dataset is the most commonly used ImageNet-1k [59], consisting of 1000 classes with a total of 1.28 million images. As for the downstream datasets, we utilize the widely adopted downstream dataset Stanford Cars [60], FGVC Aircraft [61], Oxford Flowers 102 [62], Birdsnap [63], and Oxford-IIIT Pets [64], which are fine-grained. The texture classification dataset (DTD [65]) and scene classification dataset (SUN [66]), which are significantly dissimilar to the upstream dataset, are also involved. Meanwhile, to prove the generality of our method, we also conduct experiments on Caltech-256 [67], which is similar to ImageNet-1k.

Implementation details: We choose the commonly used ResNet-50 [68] as the baseline model for all methods and follow the same pre-training procedure across all CNN-based model architectures from scratch with a batch size of 1024. We first train it for 100 epochs with an initial learning rate of 0.4, which is dropped by a factor of 10 every 30 epochs. The optimizer is SGD, and the momentum and weight decay are set as 0.9 and 1×10^{-4}, respectively. In the fine-tuning stage, we fine-tune the pre-trained network for 150 epochs with an initial learning rate of 0.01, which dropped by a factor of 10 every 50 epochs. The λ in Equation (7) is set to 0.01. All experiments are conducted on a single NVIDIA GTX 1080 Ti 11 GB GPU. In order to ensure the fairness of the comparison, for all the comparison methods, we conduct the experiments under the same setting with their published codes, and the experimental results are run three times and averaged.

4.2. Comparison with State-of-the-Arts

We compare our method on both the full-dataset setting and few-shot setting with the following state of the art: the vanilla fine-tuning model (Baseline), L2-SP [23], BSS [21], Co-Tuning [24], BYOL [50], SimCLR [51], NNCLR [52], SD [69] and NCTI [70]. We carefully retrain the first four methods under the same setting to ensure sufficient convergence of the network on the training set.

Comparison under the full-dataset setting: Table 1 lists all the quantitative results of all methods under the full-dataset setting. Owing to the effective discrepant semantic diffusion, our method achieves peak performance in almost all cases, such as 2.04% improvement on the Caltech-256 dataset. While our method records a 1.43% improvement over the baseline on the SUN dataset, it does not surpass the NCTI method. It is noteworthy that the results of NCTI are obtained from an average over a diverse array of eleven models, such as ResNet-50, ResNet-151, DenseNet-121, DenseNet-169, GoogleNet, and Inception-v3, among others. Given that our approach exclusively uses a ResNet-50 backbone, a fair comparison cannot be made. Moreover, we can find that previous methods which preserve the upstream information while under-utilizing the semantic discrepancy struggle to bring obvious improvement on the downstream datasets of mismatched semantic granularity from the upstream dataset.

For example, co-tuning considers the upstream and downstream category relationships of different datasets but neglects their discrepancies, and thus, it is difficult to yield better performance on datasets such as DTD, which are quite dissimilar to the upstream dataset. BYOL, SimCLR, and NNCLR employ the contrastive learning framework, whereas SD utilizes the synthetic images to pretrain strong general-purpose visual encoders. However, the performance of both approaches is relatively poor. In contrast, with the discrepancy

information in the form of the granularity diffusion sequence, the proposed method could achieve superior performance on both fine-grained datasets and datasets of low similarity to the upstream, such as 1.63% improvement on the DTD dataset and 1.13% improvement on the Birdsnap dataset. It is worth noting that our improvement on some datasets is less than 1%, such as Flowers 102 and Pets, which is also an obvious gain in the relatively mature transfer learning field compared to other state of the art.

Table 1. Transfer performance on eight prevalent downstream datasets under the full-dataset setting. Our model achieves almost the best performance compared to others. Bold indicates the best result.

Dataset	Method									
	Baseline	**L2-SP**	**BSS**	**Co-Tuning**	**BYOL**	**SimCLR**	**NNCLR**	**SD**	**NCTI**	**Ours**
DTD	75.18 ± 0.24	74.95 ± 0.38	75.28 ± 0.28	74.46 ± 0.24	75.50	75.70	76.70	75.90	70.40	**76.91** ± 0.23
SUN	62.83 ± 0.08	62.33 ± 0.06	62.87 ± 0.08	62.21 ± 0.16	62.20	60.60	62.50	62.50	**75.60**	64.26 ± 0.14
Caltech-256	84.28 ± 0.09	84.35 ± 0.12	84.26 ± 0.14	84.43 ± 0.11	-	-	-	-	-	**86.32** ± 0.06
Cars	90.47 ± 0.05	90.78 ± 0.06	90.73 ± 0.02	90.99 ± 0.22	67.80	49.30	67.10	57.20	64.70	**91.62** ± 0.19
Aircraft	86.65 ± 0.36	87.06 ± 0.12	87.04 ± 0.17	88.05 ± 0.23	60.60	49.80	64.10	55.30	49.60	**88.81** ± 0.21
Flowers 102	96.60 ± 0.09	96.64 ± 0.06	96.87 ± 0.18	96.82 ± 0.20	96.10	92.60	95.10	92.90	54.10	**97.44** ± 0.14
Birdsnap	72.58 ± 0.33	72.58 ± 0.18	73.32 ± 0.30	74.19 ± 0.23	57.20	42.4	61.40	-	-	**75.32** ± 0.19
Pets	93.58 ± 0.14	93.59 ± 0.16	93.54 ± 0.10	93.18 ± 0.09	90.40	84.6	91.80	88.70	92.40	**94.01** ± 0.13

Comparison under the few-shot setting: One practical application of transfer learning is to quickly adapt pre-trained information to new tasks with limited data, addressing the issue of data sparsity caused by privacy protection. Therefore, the transferring performance under the few-shot setting is a key factor in measuring the effectiveness of the transfer learning algorithm, especially in the field of AI security. Table 2 lists the results of different methods under few-shot settings. We conduct experiments on Aircraft, Cars, DTD, and SUN datasets, with 15%, 30%, and 50% of the whole data, respectively. Our method outperforms the others almost in all cases, and brings obvious gains especially with limited downstream data (15%), such as 3.75% improvement on the Cars (fine-grained) dataset and 1.79% improvement on the SUN dataset (dissimilar to upstream data), compared with suboptimal results. This proves that the proposed discrepant semantic diffusion could effectively improve the generalization of the model and is more practical, applicable and secure than the existing methods. Our method significantly reduces the training data requirements for models on new tasks, thereby lowering the potential security risks.

Table 2. Transfer performance on Aircraft, Cars, DTD, and SUN datasets under the few-shot setting. Our method outperforms the others almost in all cases, and brings obvious gains especially with limited downstream data (15%). Bold indicates the best result.

	Sampling Rates	Method				
		Baseline	**L2-SP**	**BSS**	**Co-Tuning**	**Ours**
Aircraft	15%	43.65 ± 0.25	43.02 ± 0.14	45.74 ± 0.22	43.52 ± 0.35	**48.52** ± 0.29
	30%	65.97 ± 0.29	66.08 ± 0.20	66.21 ± 0.21	66.60 ± 0.37	**67.71** ± 0.17
	50%	77.41 ± 0.34	77.39 ± 0.24	78.06 ± 0.11	77.78 ± 0.26	**79.59** ± 0.34
Cars	15%	42.27 ± 0.27	42.87 ± 0.33	43.03 ± 0.35	43.42 ± 0.37	**47.17** ± 0.18
	30%	71.55 ± 0.24	71.87 ± 0.31	71.97 ± 0.19	72.08 ± 0.21	**74.20** ± 0.16
	50%	83.44 ± 0.19	84.16 ± 0.25	84.34 ± 0.09	84.91 ± 0.21	**85.59** ± 0.22
DTD	15%	60.21 ± 0.32	61.38 ± 0.38	61.06 ± 0.39	**62.03** ± 0.24	60.44 ± 0.14
	30%	66.70 ± 0.35	66.96 ± 0.09	66.75 ± 0.26	66.62 ± 0.18	**67.16** ± 0.21
	50%	68.22 ± 0.34	69.73 ± 0.17	69.62 ± 0.31	70.15 ± 0.19	**71.73** ± 0.18
SUN	15%	45.35 ± 0.22	45.75 ± 0.15	45.02 ± 0.18	45.60 ± 0.16	**47.54** ± 0.11
	30%	53.68 ± 0.14	53.40 ± 0.14	52.86 ± 0.08	53.73 ± 0.25	**54.83** ± 0.16
	50%	57.72 ± 0.15	57.87 ± 0.15	57.53 ± 0.06	57.65 ± 0.07	**59.34** ± 0.10

4.3. Ablation Study

In this section, we verify the efficacy of each component in our method on the Aircraft dataset and report the top-1 accuracy in Figure 4. Figure 4 shows the evaluation of (I) the baseline model, (II) the baseline model with PGD, (III) the baseline model with DI2, (IV) the full model under both the full-dataset setting (100%) and few-shot setting (50%, 30%, and 15%). From Figure 4, we can observe that each contribution brings a performance gain in most cases under the full-dataset setting. On the contrary, under the few-shot setting, (II) the baseline model with PGD shows a negative performance gap. This is because without sufficient downstream data, simply adopting the semantic prior would blur the boundaries of similar categories and do harm to the downstream performance. However, when further combined with our DI2, (IV) the full model, it can achieve the best results with sufficient semantic discrepancy.

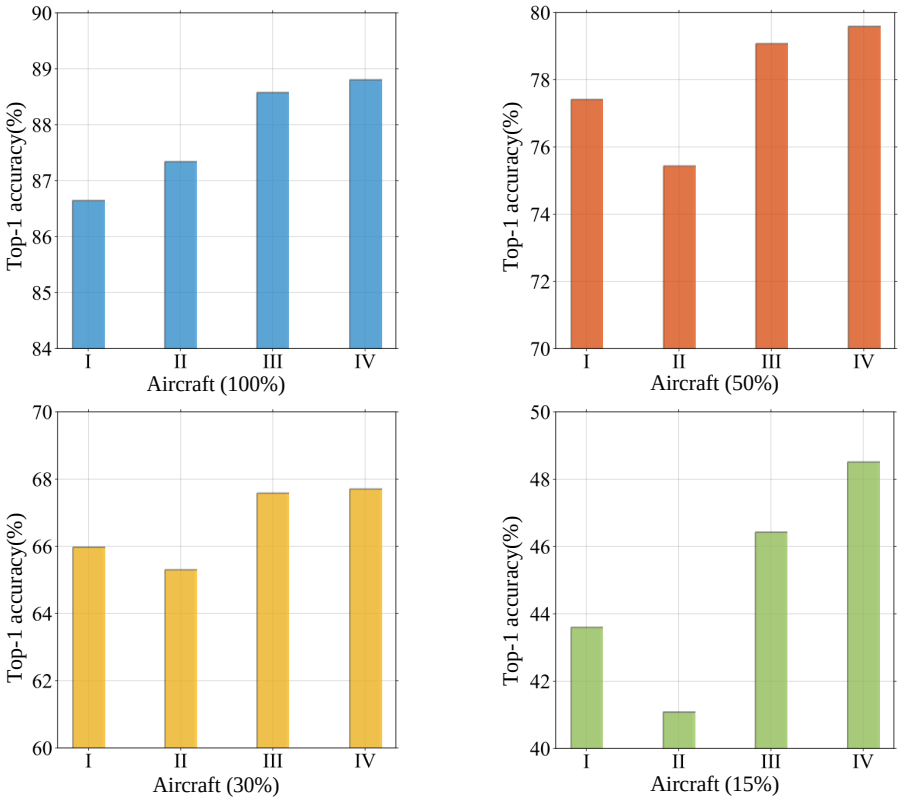

Figure 4. The ablation study on the Aircraft dataset under the full-dataset setting (100%) and few-shot setting (50%, 30%, and 15%). It shows the top-1 accuracy metrics of the (I) baseline, (II) baseline + PGD, (III) baseline + DI2, and (IV) full model, respectively.

Moreover, as shown in the results of (III) the baseline model with DI2, we can find that the proposed DI2 can achieve good performance when used alone, demonstrating its effectiveness and robustness. The baseline model performance with DI2 approaches the full model under the full-dataset setting but not when the data are insufficient. For example, in the experiment of Aircrafts with 15% of the whole data, the full model outperforms the baseline model with DI2 by about 2% as shown in Figure 4. In summary, the performance of the baseline model with DI2 is significantly inferior to that of the full model under the few-shot setting, which further proves the effectiveness of the semantic prior in PGD with sufficient semantic discrepancy involved.

4.4. Verification on Various Networks

In this section, we investigate the performance of the proposed discrepant semantic diffusion method with different backbones. We examine the transfer performance on three

common networks (ResNet-{50, 101} [68] and MobileNetV3-Large 1.0 [71]). As shown in Table 3, the proposed discrepant semantic diffusion method is effective on both a large network (ResNet101) and a light network MobileNet, indicating it could be applied with multiple network structures. This further demonstrates the generality and versatility of our method across different model architectures in different scenarios.

Table 3. Transfer performance of the proposed method with multiple backbones, which proves the versatility of the proposed discrepant semantic diffusion method with different model structures. Bold indicates the best result.

Method	Dataset			
	DTD	**SUN**	**Caltech-256**	**Cars**
Baseline (ResNet50)	75.18 ± 0.24	62.83 ± 0.08	84.28 ± 0.09	90.47 ± 0.05
Ours (ResNet50)	**76.91** ± 0.23	**64.26** ± 0.14	**86.32** ± 0.06	**91.62** ± 0.19
Baseline (ResNet101)	75.53 ± 0.10	63.23 ± 0.19	85.50 ± 0.19	91.23 ± 0.09
ours (ResNet101)	**76.59** ± 0.11	**64.73** ± 0.18	**87.17** ± 0.06	**91.65** ± 0.17
Baseline (MobileNetV3-Large)	71.64 ± 0.28	60.93 ± 0.05	82.50 ± 0.05	85.99 ± 0.10
Ours (MobileNetV3-Large)	**72.25** ± 0.06	**61.26** ± 0.06	**82.95** ± 0.12	**87.76** ± 0.13
Method	**Dataset**			
	Aircraft	**Flowers 102**	**Birdsnap**	**Pets**
Baseline (ResNet50)	86.65 ± 0.36	96.60 ± 0.09	72.58 ± 0.33	93.58 ± 0.14
Ours (ResNet50)	**88.81** ± 0.21	**97.44** ± 0.14	**75.32** ± 0.19	**94.01** ± 0.13
Baseline (ResNet101)	87.80 ± 0.31	96.79 ± 0.38	73.71 ± 0.23	93.90 ± 0.11
ours (ResNet101)	**88.86** ± 0.08	**97.43** ± 0.16	**75.41** ± 0.03	**94.22** ± 0.02
Baseline (MobileNetV3-Large)	79.83 ± 0.10	96.19 ± 0.15	68.70 ± 0.23	91.73 ± 0.15
Ours (MobileNetV3-Large)	**81.32** ± 0.26	**96.68** ± 0.08	**69.80** ± 0.03	**92.14** ± 0.08

4.5. Verification on Various Pre-Training Methods

We further prove the robustness of our DI2 method with different pre-training methods, including contrastive unsupervised pre-training MoCov2 [72], naive supervised pre-training, and our PGD. As shown in Figure 5, we can easily find that the proposed DI2 can be applied to various pre-training methods and achieve better performance. As for all the pre-training method, it is intuitive that naive supervised pre-training performs better than the unsupervised MoCov2, ranking the second. But our PGD pre-training yields the competitive results on both fine-tuning methods.

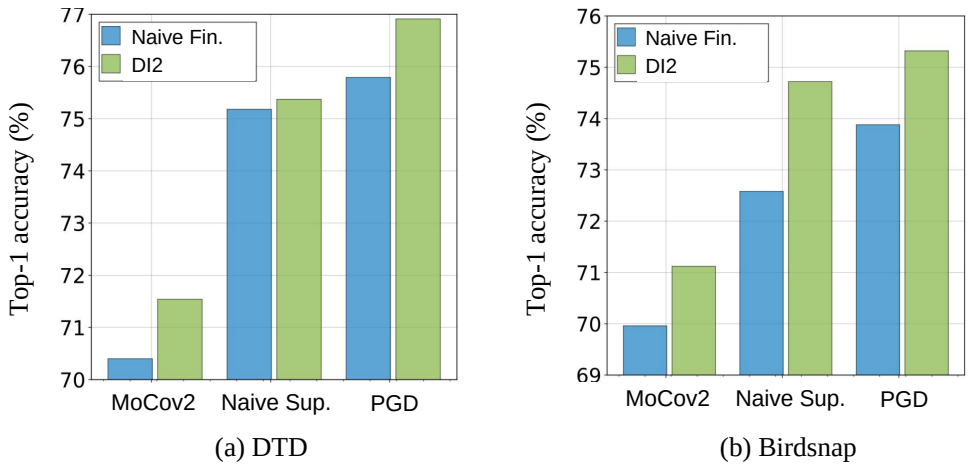

Figure 5. DI2 (green) surpasses the naive fine-tuning method (blue) when using various pre-training methods: (1) unsupervised MoCov2, (2) naive supervised pre-training, and (3) our PGD.

4.6. Hyper-Parameter Analysis

Selection of k: In this section, we investigate the effect of the hyper-parameter k in the semantic discrepancy of Section 3.3. As we mentioned, we determine k by adding

1 dissimilar class to guarantee that the granularity diffusion sequence of an arbitrary downstream category is different from the counterparts of the others. As shown in Figure 6a, we list the results of ranging the k from adding 0 ("+ 0") to adding 5 ("+ 5") dissimilar classes compared to the baseline model. We can easily find that the semantic diffusion operation can always help the model achieve better performance than the baseline model. But the "+ 0" setting, which cannot guarantee the semantic discrepancy of each of the two categories, performs worse than the other k settings. It proves that the semantic discrepancy is important for the transferring process. As k changes, our method can always achieve better and more stable performance than the baseline model. It proves that as long as the discrepant semantic diffusion can be guaranteed, the model can always achieve better transfer performance, which proves the rationality and effectiveness of our discrepancy diffusion method. Therefore, considering the training costs, we choose to add one dissimilar upstream class when determining k.

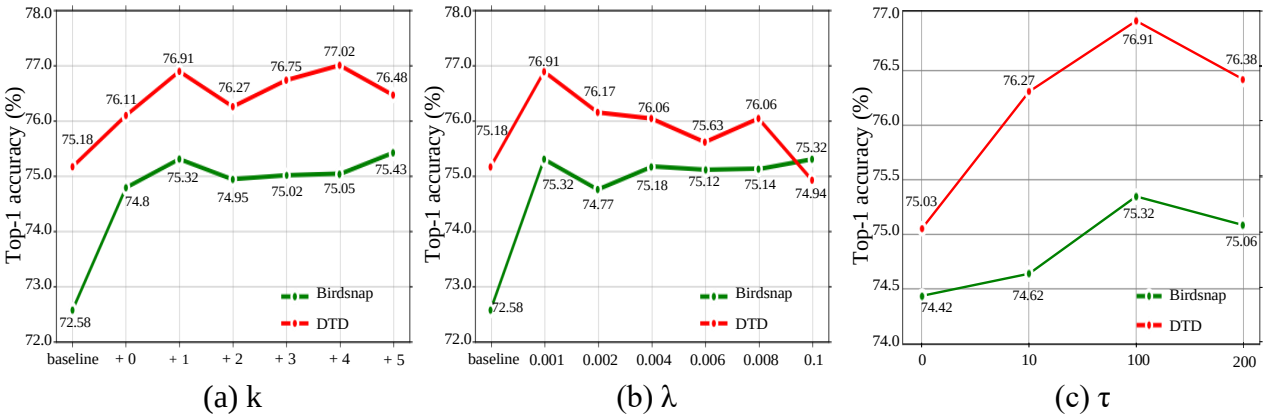

Figure 6. Analysis of hyper-parameters on SUN dataset and DTD dataset. (**a**) Analysis of the hyper-parameter k. Considering the accuracy and computational complexity, adding 1 dissimilar upstream class to the granularity diffusion sequence is chosen to guarantee semantic discrepancy. (**b**) Analysis of the hyper-parameter λ. Introducing λ always performs better than the baseline model. However, a larger λ will lead to performance degradation, so we chose 0.001 as the value of λ in the model. (**c**) Analysis of the hyper-parameter τ. A similar trend can be seen on different datasets. We choose 100 as the value of τ in the model.

Selection of λ: In this section, we investigate the sensitivity of hyper-parameter λ, which are depicted in Equation (7). As mentioned in Section 3.4, λ is introduced as a balancing factor to balance the original multi-class cross entropy loss and the proposed discrepant diffusion loss. We vary the λ to [0.001, 0.01]. The performance of the Birdsnap and DTD datasets is reported in Figure 6b. It can be easily included that with the change of the introduced hyper-parameters, the proposed methods could achieve stable performance that is better than the baseline model on a large parameter range [0.001, 0.8]. Given λ, the model can achieve similar performance on multiple datasets. Therefore, it can be proved that our method has better robustness to the introduced hyper-parameters. It is worth noting that, for the DTD dataset, a bigger λ, such as $\lambda = 1$, leads to poor performance. It is because the DTD dataset has less similarity to the upstream dataset. Therefore, a particularly large λ can affect the original classification loss, resulting in poor classification performance.

Selection of τ: In this section, we investigate the effect of the hyper-parameter τ in Equation (3) on the Birdsnap and DTD datasets. The hyper-parameter could control the penalty degree for the ground-truth class, the similar classes of the ground-truth classes, and the others. From Figure 6c, we can observe that, with the increasing of τ, the top-1 accuracy first rises, reaching 76.91% when $\tau = 100$, and then drops. It can be concluded that a larger τ would yield competitive performance since it could model the reasonable class relation and enable the semantic diffusion ability. However, when τ becomes too large,

it will affect the semantic discrimination. Similar trends could be seen on both datasets. Therefore, we set $\tau = 100$ under all the settings.

4.7. Visualization

Figure 7 shows the t-SNE visualizations of embeddings of “icefloe" and “snowfield" in the SUN dataset. We show the embeddings extracted from (a) the pretrained model, (b) the baseline fine-tuned model and (c) the model with our discrepant semantic diffusion, respectively. As Figure 7 depicts, the pretrained model shows collapsed classification due to the mismatched downstream–upstream semantic granularity, and thus two categories are mixed in the embedding space and cannot be distinguished. And rigidly preserving the upstream information while fine-tuning with the downstream data cannot well address this problem, without considering the discrepant downstream–upstream mapping to ensure the downstream classification. In contrast, our method could obviously alleviate the collapsed classification problem, benefiting from the proposed discrepant semantic diffusion and thus achieve superior performance.

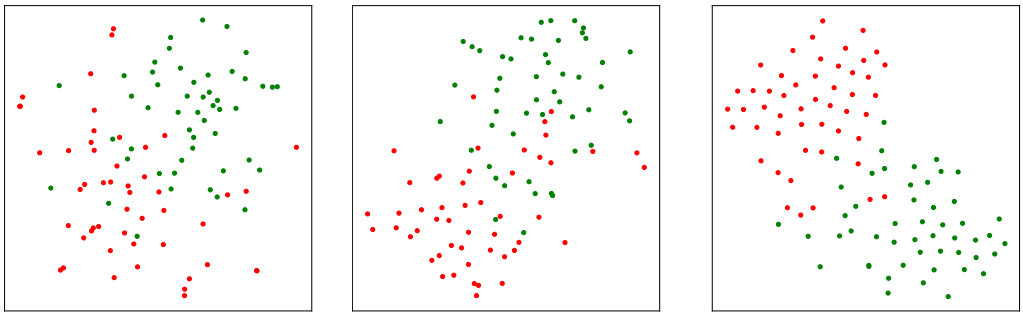

Figure 7. t-SNE visualization of embeddings of icefloe and snowfield in the SUN dataset. We show the embeddings extracted from (**a**) the pretrained model, (**b**) the baseline model fine-tuned on the SUN dataset and (**c**) the model with our discrepant semantic diffusion. The baseline model cannot address the indistinguishable embeddings and the collapsed classification, while ours could alleviate these problem and achieve superior performance.

5. Conclusions

In this paper, we propose a discrepant semantic diffusion method for transfer learning to handle the downstream collapsed classification problem and further boost the transfer learning robustness. The proposed framework consists of Prior-Guided Diffusion (PGD) for pre-training and Discrepant Diffusion (DI2) for fine-tuning. The PGD utilizes a semantic prior to model the valuable semantic connection, empowering the pre-trained model with the semantic diffusion ability. As a result, the pre-trained model becomes more generalized, paving the way for enhanced downstream discrimination. The DI2 in the fine-tuning stage models the diffusive downstream–upstream connection constrained by the semantic discrepancy, effectively aligning the shared downstream–upstream knowledge. Extensive experiments prove the superiority of our method for various downstream tasks, especially for the fine-grained datasets or datasets dissimilar to the upstream data (e.g., 3.75% and 1.79% improvements for the Cars and SUN datasets, respectively).

Author Contributions: Conceptualization, Y.G., S.B., X.Z., R.G., Y.W. and Y.M.; methodology, Y.G., S.B., R.G. and Y.M.; software, X.Z., Y.W. and R.G.; validation, Y.G., S.B. and Y.M.; formal analysis, Y.G., S.B. and X.Z.; investigation, Y.G., S.B., X.Z., R.G. and Y.M.; resources, R.G., Y.W. and X.Z.; data curation, S.B.; writing—original draft preparation, Y.G. and S.B.; writing—review and editing, Y.G., X.Z., R.G., Y.W. and Y.M.; visualization, Y.G. and S.B.; supervision, Y.M.; project administration, Y.M. and R.G.; funding acquisition, Y.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grant 62206010 and 62022009.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: We thank the State Key Lab of Software Development Environment for providing the experimental environment and equipment.

Conflicts of Interest: Authors S.B. and R.G. were employed by the company SenseTime. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
PGD	Prior-Guided Diffusion
DI2	Discrepant Diffusion
SGD	Stochastic Gradient Descent

References

1. Chakraborty, C.; Nagarajan, S.M.; Devarajan, G.G.; Ramana, T.; Mohanty, R. Intelligent AI-based Healthcare Cyber Security System using Multi-Source Transfer Learning Method. *ACM Trans. Sens. Netw.* **2023**. [CrossRef]
2. Yılmaz, S.; Aydogan, E.; Sen, S. A transfer learning approach for securing resource-constrained iot devices. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 4405–4418. [CrossRef]
3. Singla, A.; Bertino, E.; Verma, D. Overcoming the lack of labeled data: Training intrusion detection models using transfer learning. In Proceedings of the 2019 IEEE International Conference on Smart Computing (SMARTCOMP), Washington, DC, USA, 12–15 June 2019; pp. 69–74.
4. Pan, W.; Xiang, E.; Liu, N.; Yang, Q. Transfer learning in collaborative filtering for sparsity reduction. In Proceedings of the AAAI Conference on Artificial Intelligence, Atlanta, GA, USA, 11–15 July 2010; Volume 24, pp. 230–235.
5. Rezaei, S.; Liu, X. A target-agnostic attack on deep models: Exploiting security vulnerabilities of transfer learning. *arXiv* **2019**, arXiv:1904.04334.
6. Zhou, Z.; Hu, S.; Zhao, R.; Wang, Q.; Zhang, L.Y.; Hou, J.; Jin, H. Downstream-agnostic adversarial examples. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 4345–4355.
7. Wang, B.; Yao, Y.; Viswanath, B.; Zheng, H.; Zhao, B.Y. With great training comes great vulnerability: Practical attacks against transfer learning. In Proceedings of the 27th USENIX Security Symposium (USENIX Security 18), Baltimore, MD, USA, 15–17 August 2018; pp. 1281–1297.
8. Park, J.; Low, C.Y.; Beng Jin Teoh, A. Divergent Angular Representation for Open Set Image Recognition. *IEEE Trans. Image Process.* **2021**, *31*, 176–189. [CrossRef] [PubMed]
9. Tian, H.; Qu, S.; Payeur, P. A Prototypical Knowledge Oriented Adaptation Framework for Semantic Segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 149–163. [CrossRef] [PubMed]
10. Chen, S.B.; Wei, Q.S.; Wang, W.Z.; Tang, J.; Luo, B.; Wang, Z.Y. Remote Sensing Scene Classification via Multi-Branch Local Attention Network. *IEEE Trans. Image Process.* **2022**, *31*, 99–109. [CrossRef] [PubMed]
11. Zhu, Y.; Chen, Y.; Lu, Z.; Pan, S.; Xue, G.R.; Yu, Y.; Yang, Q. Heterogeneous transfer learning for image classification. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 7–11 August 2011; Volume 25, pp. 1304–1309.
12. Shi, X.; Liu, Q.; Fan, W.; Philip, S.Y.; Zhu, R. Transfer learning on heterogenous feature spaces via spectral transformation. In Proceedings of the 2010 IEEE International Conference on Data Mining, Sydney, Australia, 13–17 December 2010; pp. 1049–1054.
13. Hussain, M.; Fiza, M.; Khalil, A.; Siyal, A.A.; Dharejo, F.A.; Hyder, W.; Guzzo, A.; Krichen, M.; Fortino, G. Transfer learning-based quantized deep learning models for nail melanoma classification. *Neural Comput. Appl.* **2023**, *35*, 22163–22178. [CrossRef]
14. Boulouard, Z.; Ouaissa, M.; Ouaissa, M.; Krichen, M.; Almutiq, M.; Gasmi, K. Detecting Hateful and Offensive Speech in Arabic Social Media Using Transfer Learning. *Appl. Sci.* **2022**, *12*, 12823. [CrossRef]
15. Zou, J.; Guo, W.; Wang, F. A Study on Pavement Classification and Recognition Based on VGGNet-16 Transfer Learning. *Electronics* **2023**, *12*, 3370. [CrossRef]
16. Zhou, F.; Hu, S.; Wan, X.; Lu, Z.; Wu, J. Diplin: A Disease Risk Prediction Model Based on EfficientNetV2 and Transfer Learning Applied to Nursing Homes. *Electronics* **2023**, *12*, 2581. [CrossRef]
17. Nouman Noor, M.; Nazir, M.; Khan, S.A.; Song, O.Y.; Ashraf, I. Efficient gastrointestinal disease classification using pretrained deep convolutional neural network. *Electronics* **2023**, *12*, 1557. [CrossRef]

18. Gao, L.; Zhang, X.; Yang, T.; Wang, B.; Li, J. The Application of ResNet-34 Model Integrating Transfer Learning in the Recognition and Classification of Overseas Chinese Frescoes. *Electronics* **2023**, *12*, 3677. [CrossRef]
19. Yu, Z.; Shen, D.; Jin, Z.; Huang, J.; Cai, D.; Hua, X.S. Progressive Transfer Learning. *IEEE Trans. Image Process.* **2022**, *31*, 1340–1348. [CrossRef] [PubMed]
20. Li, X.; Xiong, H.; Wang, H.; Rao, Y.; Liu, L.; Huan, J. DELTA: DEep Learning Transfer using Feature Map with Attention for Convolutional Networks. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
21. Chen, X.; Wang, S.; Fu, B.; Long, M.; Wang, J. Catastrophic Forgetting Meets Negative Transfer: Batch Spectral Shrinkage for Safe Transfer Learning. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 1906–1916.
22. Guo, Y.; Shi, H.; Kumar, A.; Grauman, K.; Rosing, T.; Feris, R. SpotTune: Transfer Learning Through Adaptive Fine-Tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
23. Li, X.; Grandvalet, Y.; Davoine, F. Explicit Inductive Bias for Transfer Learning with Convolutional Networks. In Proceedings of the 35th International Conference on Machine Learning, (ICML), Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 2830–2839.
24. You, K.; Kou, Z.; Long, M.; Wang, J. Co-tuning for transfer learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17236–17246.
25. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27, pp. 3320–3328.
26. Zhao, N.; Wu, Z.; Lau, R.W.H.; Lin, S. What makes instance discrimination good for transfer learning? In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
27. Salman, H.; Ilyas, A.; Engstrom, L.; Kapoor, A.; Madry, A. Do Adversarially Robust ImageNet Models Transfer Better? In Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS), Virtual, 6–12 December 2020.
28. Abd El-Rady, A.; Osama, H.; Sadik, R.; El Badwy, H. Network Intrusion Detection CNN Model for Realistic Network Attacks Based on Network Traffic Classification. In Proceedings of the 2023 40th National Radio Science Conference (NRSC), Giza, Egypt, 30 May–1 June 2023; Volume 1, pp. 167–178.
29. Alabsi, B.A.; Anbar, M.; Rihan, S.D.A. CNN-CNN: Dual Convolutional Neural Network Approach for Feature Selection and Attack Detection on Internet of Things Networks. *Sensors* **2023**, *23*, 6507. [CrossRef] [PubMed]
30. Li, X.; Wu, J.; Sun, Z.; Ma, Z.; Cao, J.; Xue, J.H. BSNet: Bi-Similarity Network for Few-shot Fine-grained Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 1318–1331. [CrossRef] [PubMed]
31. Ding, Y.; Ma, Z.; Wen, S.; Xie, J.; Chang, D.; Si, Z.; Wu, M.; Ling, H. AP-CNN: Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-Grained Visual Classification. *IEEE Trans. Image Process.* **2021**, *30*, 2826–2836. [CrossRef]
32. Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; Darrell, T. DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proceedings of the 31th International Conference on Machine Learning, (ICML), Beijing, China, 21–26 June 2014; Volume 32, pp. 647–655.
33. Jing, L.; Chen, Y.; Tian, Y. Coarse-to-fine semantic segmentation from image-level labels. *IEEE Trans. Image Process.* **2019**, *29*, 225–236. [CrossRef]
34. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI).* **2017**, 40, 834–848. [CrossRef]
35. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the IEEE international conference on computer vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969
36. Ma, W.Y.; Manjunath, B. EdgeFlow: A technique for boundary detection and image segmentation. *IEEE Trans. Image Process.* **2000**, *9*, 1375–1388.
37. Fang, F.; Li, L.; Zhu, H.; Lim, J.H. Combining Faster R-CNN and Model-Driven Clustering for Elongated Object Detection. *IEEE Trans. Image Process.* **2019**, *29*, 2052–2065. [CrossRef] [PubMed]
38. Wang, H.; Wang, Q.; Zhang, H.; Hu, Q.; Zuo, W. CrabNet: Fully Task-Specific Feature Learning for One-Stage Object Detection. *IEEE Trans. Image Process.* **2022**, *31*, 2962–2974. [CrossRef] [PubMed]
39. Li, J.; Li, J.; Zhu, L.; Xiang, X.; Huang, T.; Tian, Y. Asynchronous Spatio-Temporal Memory Network for Continuous Event-Based Object Detection. *IEEE Trans. Image Process.* **2022**, *31*, 2975–2987. [CrossRef] [PubMed]
40. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/accuracy trade-offs for modern convolutional object detectors. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7310–7311.
41. Liang, Y.; Pan, Y.; Lai, H.; Liu, W.; Yin, J. Deep Listwise Triplet Hashing for Fine-Grained Image Retrieval. *IEEE Trans. Image Process.* **2022**, *31*, 949–961. [CrossRef] [PubMed]
42. Yang, H.; Yan, D.; Zhang, L.; Sun, Y.; Li, D.; Maybank, S.J. Feedback Graph Convolutional Network for Skeleton-Based Action Recognition. *IEEE Trans. Image Process.* **2022**, *31*, 164–175. [CrossRef] [PubMed]

43. Li, H.; Jiang, X.; Guan, B.; Tan, R.R.M.; Wang, R.; Thalmann, N.M. Joint Feature Optimization and Fusion for Compressed Action Recognition. *IEEE Trans. Image Process.* **2021**, *30*, 7926–7937. [CrossRef] [PubMed]
44. Liu, F.; Shen, C.; Lin, G. Deep Convolutional Neural Fields for Depth Estimation from a Single Image. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5162–5170.
45. Phoo, C.P.; Hariharan, B. Self-training for Few-shot Transfer Across Extreme Task Differences. *arXiv* **2020**, arXiv:2010.07734.
46. Mormont, R.; Geurts, P.; Marée, R. Comparison of Deep Transfer Learning Strategies for Digital Pathology. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, UT, USA, 18–22 June 2018; pp. 2343–234309.
47. Kruggel, F. A Simple Measure for Acuity in Medical Images. *IEEE Trans. Image Process.* **2018**, *27*, 5225–5233. [CrossRef]
48. Pan, S.J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]
49. Kolesnikov, A.; Beyer, L.; Zhai, X.; Puigcerver, J.; Yung, J.; Gelly, S.; Houlsby, N. Large Scale Learning of General Visual Representations for Transfer. *arXiv* **2019**, arXiv:1912.11370.
50. Grill, J.B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 21271–21284.
51. Chen, T.; Kornblith, S.; Norouzi, M.; Hinton, G. A simple framework for contrastive learning of visual representations. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 1597–1607.
52. Dwibedi, D.; Aytar, Y.; Tompson, J.; Sermanet, P.; Zisserman, A. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 9588–9597.
53. Sariyildiz, M.B.; Kalantidis, Y.; Alahari, K.; Larlus, D. No reason for no supervision: Improved generalization in supervised models. In Proceedings of the ICLR 2023-International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023; pp. 1–26.
54. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.C.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [CrossRef] [PubMed]
55. Kornblith, S.; Shlens, J.; Le, Q.V. Do Better ImageNet Models Transfer Better? In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2661–2671.
56. Li, H.; Chaudhari, P.; Yang, H.; Lam, M.; Ravichandran, A.; Bhotika, R.; Soatto, S. Rethinking the Hyperparameters for Fine-tuning. *arXiv* **2020**, arXiv:2002.11770.
57. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event, Austria, 3–7 May 2021.
58. Steiner, A.; Kolesnikov, A.; Zhai, X.; Wightman, R.; Uszkoreit, J.; Beyer, L. How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers. *arXiv* **2021**, arXiv:2106.10270.
59. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
60. Krause, J.; Deng, J.; Stark, M.; Fei-Fei, L. Collecting a Large-Scale Dataset of Fine-Grained Cars. In Second Workshop on Fine-Grained Visual Categorization. 2013. Available online: https://ai.stanford.edu/~jkrause/papers/fgvc13.pdf (accessed on 12 November 2023).
61. Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.B.; Vedaldi, A. Fine-Grained Visual Classification of Aircraft. *arXiv* **2013**, arXiv:1306.5151.
62. Nilsback, M.E.; Zisserman, A. Automated Flower Classification over a Large Number of Classes. In Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing, Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
63. Berg, T.; Liu, J.; Lee, S.W.; Alexander, M.L.; Jacobs, D.W.; Belhumeur, P.N. Birdsnap: Large-Scale Fine-Grained Visual Categorization of Birds. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
64. Parkhi, O.M.; Vedaldi, A.; Zisserman, A.; Jawahar, C.V. Cats and dogs. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3498–3505.
65. Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; Vedaldi, A. Describing Textures in the Wild. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 3606–3613
66. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. SUN database: Large-scale scene recognition from abbey to zoo. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
67. Griffin, G.; Holub, A.; Perona, P. *Caltech-256 object category dataset*; California Institute of Technology: Pasadena, CA, USA, 2007.
68. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
69. Sariyildiz, M.B.; Alahari, K.; Larlus, D.; Kalantidis, Y. Fake it till you make it: Learning transferable representations from synthetic ImageNet clones. In Proceedings of the CVPR 2023–IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.

70. Wang, Z.; Luo, Y.; Zheng, L.; Huang, Z.; Baktashmotlagh, M. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 5549–5558.
71. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
72. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved Baselines with Momentum Contrastive Learning. *arXiv* **2020**, arXiv:2003.04297.