applied sciences

MDPI

Article

Patch-Level Consistency Regularization in Self-Supervised Transfer Learning for Fine-Grained Image Recognition

Yejin Lee [1,†], **Suho Lee** [1] **and Sangheum Hwang** [1,2,3,*]

[1] Department of Data Science, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea; yejin2@ds.seoultech.ac.kr (Y.L.); swlee@ds.seoultech.ac.kr (S.L.)
[2] Department of Industrial & Information Systems Engineering, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
[3] Research Center for Electrical and Information Technology, Seoul National University of Science and Technology, Seoul 01811, Republic of Korea
* Correspondence: shwang@seoultech.ac.kr
† Current address: NAVER Z Corporation, Seoul 13529, Republic of Korea

Abstract: Fine-grained image recognition aims to classify fine subcategories belonging to the same parent category, such as vehicle model or bird species classification. This is an inherently challenging task because a classifier must capture subtle interclass differences under large intraclass variances. Most previous approaches are based on supervised learning, which requires a large-scale labeled dataset. However, such large-scale annotated datasets for fine-grained image recognition are difficult to collect because they generally require domain expertise during the labeling process. In this study, we propose a self-supervised transfer learning method based on Vision Transformer (ViT) to learn finer representations without human annotations. Interestingly, it is observed that existing self-supervised learning methods using ViT (e.g., DINO) show poor patch-level semantic consistency, which may be detrimental to learning finer representations. Motivated by this observation, we propose a consistency loss function that encourages patch embeddings of the overlapping area between two augmented views to be similar to each other during self-supervised learning on fine-grained datasets. In addition, we explore effective transfer learning strategies to fully leverage existing self-supervised models trained on large-scale labeled datasets. Contrary to the previous literature, our findings indicate that training only the last block of ViT is effective for self-supervised transfer learning. We demonstrate the effectiveness of our proposed approach through extensive experiments using six fine-grained image classification benchmark datasets, including FGVC Aircraft, CUB-200-2011, Food-101, Oxford 102 Flowers, Stanford Cars, and Stanford Dogs. Under the linear evaluation protocol, our method achieves an average accuracy of 78.5%, outperforming the existing transfer learning method, which yields 77.2%.

Keywords: self-supervised learning; fine-grained image recognition; transfer learning; Vision Transformer

check for updates

Citation: Lee, Y.; Lee, S.; Hwang, S. Patch-Level Consistency Regularization in Self-Supervised Transfer Learning for Fine-Grained Image Recognition. *Appl. Sci.* **2023**, *13*, 10493. https://doi.org/10.3390/app131810493

Academic Editors: Dongpo Xu, Huisheng Zhang and Jie Yang

Received: 8 August 2023
Revised: 16 September 2023
Accepted: 17 September 2023
Published: 20 September 2023

1. Introduction

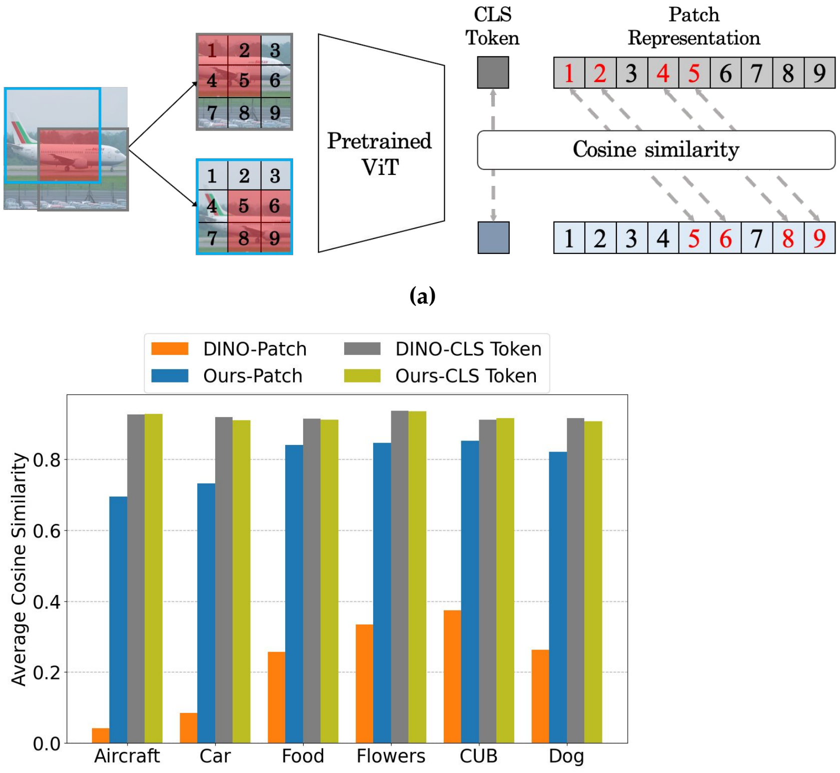
Self-supervised learning (SSL) has recently made significant progress in various fields, including computer vision [1,2], natural language processing [3], and graph representation learning [4]. SSL aims to learn generic feature representations by encouraging a model to solve auxiliary tasks that arise from the inherent properties of the data themselves. The predominant SSL approaches in computer vision seek to maximize the agreement between different views of an image [1,2,5,6]. Thanks to its strong ability to learn visual representations in the absence of human annotations, SSL has emerged as a promising strategy to reduce the reliance on large-scale labeled datasets. Remarkably, SSL-learned visual representations perform better than supervised learning [1] when transferred to several downstream vision tasks.

SSL can be especially useful for tasks requiring heavy annotation costs, such as fine-grained image recognition [7], because it aims to learn discriminative representations without using human annotation. However, there are several limitations to the application of current SSL methods for fine-grained image recognition tasks [8–10]. In contrast with ordinary computer vision tasks, fine-grained images share many visual characteristics across their classes. Therefore, a model should learn finer representations that capture subtle differences among classes. A model trained to classify bird species, for instance, should be able to learn local patterns, such as beak length, wing shape, and tail color. However, it is known that existing SSL methods tend to focus on background pixels and low-level features (e.g., texture and color) [11], which might be detrimental to learning finer representations for foreground objects. In addition, SSL with fine-grained datasets may not be as effective as anticipated because the fine-grained image dataset is relatively small in scale, and SSL is known to benefit from large-scale datasets.

In this study, we focus on SSL for fine-grained image recognition tasks based on the Vision Transformer (ViT) [12] architecture, which has recently shown remarkable performance in image recognition tasks. In the ViT architecture, an image is divided into multiple patches, and these patches are converted into a sequence of linear embeddings. Additionally, a learnable embedding vector, named [CLS] token, is prepended to the embedded patches to form an input sequence. The final output feature that corresponds to the [CLS] token serves as the image representation, which is then passed to the classification head for prediction. ViT learns visual representations using self-attention between image patches rather than convolution operations, enabling the efficient encoding of patch-level representations. Nevertheless, most existing SSL methods force image-level representations to be invariant to different image augmentations and discard the final patch-level representations. However, such information may not be adequately encoded in image-level representation without explicit regularization because class-discriminative patterns for fine-grained images are likely to appear in the local area.

To consolidate our motivation, we empirically examined the consistency of the patch representation (i.e., the final output feature of each patch) from ViT pretrained by DINO [1] on ImageNet, as shown in Figure 1a. To this end, the two differently cropped views from an image were fed into the model, and the cosine similarity between the patch representations corresponding to the same patches in each view was measured. If local semantic information is well-encoded in the patch representation, the last features of the same patches will be consistent even with different adjacent patches and thus will show high similarity. However, DINO, a state-of-the-art SSL method with ViT, does not satisfy this consistency, as shown in Figure 1b. On various public fine-grained visual classification (FGVC) datasets, DINO shows strong consistency between image-level representations (i.e., the final feature of the [CLS] token) but shows poor patch-level consistency. We argue that a model can better attend to such local information to produce image-level representations if the semantic information of each patch is properly encoded in its patch-level representations and eventually can learn finer representations that are beneficial to fine-grained image recognition.

Motivated by the above discussion, we propose a novel SSL framework that considers patch-level semantic information to enhance patch-level consistency. Specifically, if an overlapping region exists between two different views of an image, the proposed consistency loss encourages the embeddings of the overlapping patches to be similar. For SSL with fine-grained images, the ImageNet pretrained model (through SSL) is transferred to the FGVC task to leverage useful representations from large-scale datasets [13]. We found that constraining the number of learnable parameters during SSL is helpful for FGVC downstream tasks. Our extensive experimental results on various FGVC datasets demonstrated that the proposed consistency loss can further improve the quality of fine-grained visual representations.



(a)

(b)

Figure 1. (**a**) Procedure of evaluating patch-level consistency. (**b**) Average cosine similarity of image and patch-level representations on various FGVC datasets. DINO shows low cosine similarity between patch representations corresponding to the overlapping area, which indicates poor patch-level consistency.

The contributions of this study can be summarized as follows:

- We explore the effective transfer learning strategies of self-supervised pretrained representations for SSL with small-scale FGVC datasets. Specifically, it is demonstrated that high-quality representations can be attained when only the last block of ViT is updated during transfer learning.
- We propose a novel consistency loss that considers patch-level semantic information to learn fine-grained visual representations with SSL. As an auxiliary loss function, it encourages a model to produce consistent representations for the overlapping patches in augmented views.
- The effectiveness of the proposed method is demonstrated on six different FGVC datasets, including CUB200-2011 [14], Stanford Car [15], FGVC Aircraft [16], etc. It is verified quantitatively and qualitatively that our method is effective in learning fine-grained representations via SSL. Contrary to existing SSL methods, we show that the proposed loss encourages a model to learn semantically consistent patch embedding.

2. Related Works

2.1. *Self-Supervised Learning*

Self-supervised learning (SSL) aims to learn useful representations from data, without human annotation. Contrastive learning [2,5,17,18] has gained significant attention for SSL owing to its superior performance. Contrastive learning aims to maximize the agreement between different views from an image (i.e., a positive pair) while repelling those from other images (i.e., negative pairs) in the feature space. However, these approaches usually incur substantial memory costs because they require numerous negative samples during training, such as a large batch size [2] or a large memory bank [5]. Several alternatives have been proposed for effectively learning visual representations without using negative samples to address this problem. Grill et al. [19] proposed training an online network by predicting the output of a momentum encoder with a stop-gradient operator. Zbontar et al. [6] proposed optimizing the empirical cross-correlation matrix, which was obtained from a batch of feature embeddings, to be similar to the identity matrix.

Recently, several attempts have been made to apply SSL to ViT [1,20,21]. For example, Caron et al. [1] proposed an SSL framework based on the ViT architecture named self-distillation with no labels (DINO). DINO adopts knowledge distillation within an SSL framework. Specifically, the momentum encoder is treated as a teacher network and the student network is trained to match the output distribution of the teacher network, by minimizing the cross-entropy. However, the patch-level representation may not contain meaningful information, as shown in Figure 1b, because DINO uses only the last feature of the [CLS] token for training.

2.2. Fine-Grained Visual Classification

Fine-grained visual classification (FGVC) is a computer vision task focused on distinguishing between objects that are visually similar and belong to closely related classes. In FGVC, a model should be able to capture subtle interclass differences under large intraclass variance, which presents intrinsic challenges. Most of the existing FGVC approaches fall into two categories: object-part-based methods and attention-based methods [22].

Earlier works in object-part-based methods use detection or segmentation techniques to locate important regions, and then the localized information is used as a discriminative partial-level representation [23–25]. While these methods have demonstrated their effectiveness, they require bounding-box and segmentation annotations, resulting in a significant effort to obtain supervised annotations.

In contrast, attention-based methods [26–28] use attention mechanisms to improve feature learning and identify object details, thus eliminating the need for dense annotations. RA-CNN [26] iteratively generates region attention maps in a coarse-to-fine manner, using previous predictions as a reference. PCA-Net [27] uses image pairs of the same category to compute attention between feature maps to capture the common discriminative features. With the great success of ViT in computer vision, there have been several attempts to extend the use of ViT in FGVC, such as TransFG [29], SIM-Trans [30], and AFTrans [31]. Similarly, these approaches utilize self-attention maps to enhance feature learning and capture object details. While these studies have achieved considerable success, the challenge posed by high annotation costs remains. In this study, we seek to explore techniques that enable the learning of finer representations without relying on label information.

2.3. Transfer Learning

Transfer learning is a popular approach that aims to transfer pretrained knowledge to various domains and tasks [32]. A widely used strategy is to fine-tune a large-scale pretrained model (e.g., ImageNet) to a downstream task or dataset. The effect of transfer learning can be interpreted as an expert in a specific field quickly adapting to similar fields. Accordingly, it can be expected to achieve high performance even with a small amount of training data [33,34]. Thus, we considered transferring the ImageNet pretrained model via SSL to FGVC tasks by setting the pretrained weight as the initial parameter.

Although numerous methods for effective transfer learning have been proposed, most of them consider supervised learning for downstream tasks [35,36]. Contrary to previous approaches, we study SSL for downstream tasks because our goal is to learn finer representations without label information. Similarly, hierarchical pretraining (HPT) [13] transfers SSL-pretrained models by conducting SSL once again on the downstream dataset. While HPT is similar to our proposed method in assuming SSL for downstream tasks, they focus on CNN-based architectures and have shown that only updating normalization layers, such as batch normalization [37] during transfer learning is effective. In this study, we focus on the ViT architecture and suggest a more effective strategy that only trains the last block of ViT during transfer learning.

3. Methods

3.1. Transfer Learning

SSL does not require a human-annotated dataset in the training process but is based on the fact that unlabeled data are rich in themselves [2]. Furthermore, ViT has a lower locality inductive bias than CNNs; therefore, it is difficult to train with a small dataset [38]. Training a self-supervised model based on ViT from scratch is challenging because FGVC datasets are generally small in scale.

In this study, we address this issue by transfer learning from ImageNet pretrained models. Specifically, the initial parameters of ViT are set to the pretrained weights with DINO and then optimized in a self-supervised manner on FGVC datasets. Meanwhile, FGVC datasets are typically small scale, where SSL is known to benefit from training on large-scale unlabeled data. In this situation, updating all parameters during transfer learning may lead to suboptimal performance. Previous literature has shown that updating only the normalization layer when transferring a self-supervised pretrained model with SSL can be an effective strategy [13]. However, these results are mainly obtained with CNN models, and it remains unclear for ViT due to the architectural difference. Hence, we conducted experiments by hierarchically setting the learnable parameters to explore effective transfer learning strategies. We evaluated the scenarios of learning all parameters of ViT, just the normalization layer following [13], or the last block.

3.2. Consistency Loss

We propose regularizing patch-level features explicitly with consistency representation learning, as shown in Figure 2. First, we create augmented views of an input image x that includes local and global views. Each augmented view is an input to the teacher and student networks, which have the same architecture but different parameters. The final output of the [CLS] token contains the overall semantic information of the image, and the patch embedding vectors focus on a more local region (i.e., each patch). The outputs from each [CLS] token should be similar in the embedding space since the augmented views generated from a single image share the same semantic information. To this end, we consider the pretraining objective in DINO, a noncontrastive SSL framework:

$$\mathcal{L}_{[CLS]} = -P\left(g_{[CLS]}\right)^{\top} \log P\left(l_{[CLS]}\right), \tag{1}$$

where $g_{[CLS]}$ and $l_{[CLS]}$ are image-level representations, the outputs of the projection head h, and P denotes a softmax transformation. By minimizing the cross-entropy loss, the model learns to match the image-level representations of the augmented views.

To further enhance the quality of the representations, we propose a consistency loss $\mathcal{L}_{con}$ for patch embedding vectors to learn internal fine-grained structures within an image:

$$\mathcal{L}_{con} = -\frac{1}{N} \sum_{i=1}^{N=H\times W} P\left(g^{i}_{patch}\right)^{\top} \log P\left(l^{i}_{patch}\right), \tag{2}$$

where g_{patch} and l_{patch} are patch-level representations associated with overlapping regions of the two augmented views. As shown in Figure 2, the region of interest (RoI) align layer is applied to the corresponding overlapping features to match the spatial dimension of overlapping representations from two different global and local views. H and W are the height and width of the output feature map from the RoI align layer, respectively. The model is capable of learning local fine-grained features by encouraging the distribution of patch-level features corresponding to the same area (i.e., patches representing the same semantics) to be similar. As shown in Figure 1b, the proposed consistency loss significantly contributes to the generation of consistent representations for the overlapping patch tokens and the [CLS] token. The output feature of a specific patch depends on the surrounding patches because ViT is a self-attention-based architecture. However, overlapping patches still have the same semantic information because they originate from a single image. Therefore, it is reasonable to produce similar representations for these patches. Note that the consistency loss function is not calculated if there is no overlapping region between augmented views.

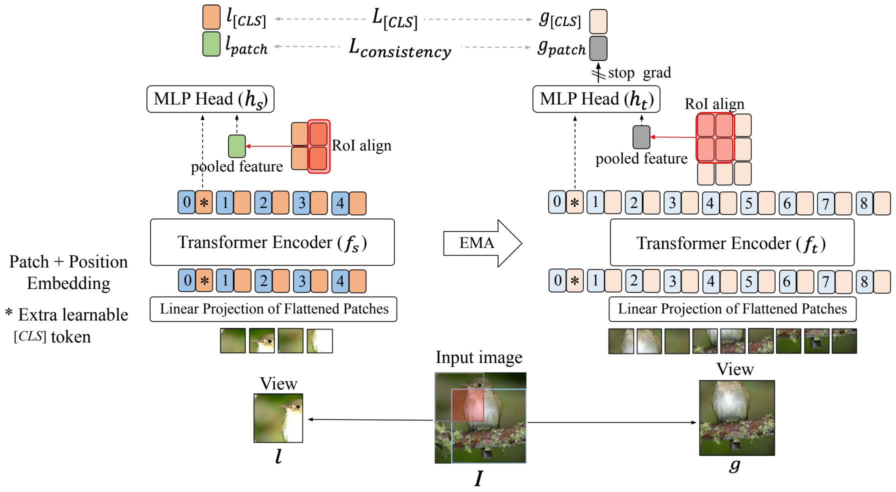


Figure 2. Framework overview of the proposed consistency loss. The '∗' symbol represents an extra learnable [CLS] token. We first find the intersection region between the global view (g) and local view (l) and generate a corresponding feature map by RoI align pooling. Through the consistency loss, the RoI feature vectors of the two views are to be similar feature vectors.

The overall SSL objective is $\mathcal{L} = \mathcal{L}_{[CLS]} + \lambda \mathcal{L}_{con}$, which learns both global- and local-level representations. We applied a stop-gradient operator to the teacher network, updated it with an exponential moving average (EMA) of the student parameters, and set the parameter $\lambda = 0.5$ for all experiments.

4. Experiment

We performed two sets of experiments to demonstrate the effectiveness of the proposed method for fine-grained visual classification. It was confirmed that applying consistency loss during SSL yields better representations on the six FGVC benchmark datasets. We also observed that freezing the lower blocks of the pretrained SSL model on ImageNet is effective for transfer learning with small-scale FGVC datasets. Furthermore, the proposed consistency loss was applied to SSL with ImageNet, a representative large-scale coarse-grained dataset. We show that our method is advantageous for learning fine-grained representations through qualitative and quantitative evaluations even though ImageNet classification performance slightly decreases compared with DINO.

4.1. Dataset

We conducted experiments on six widely used datasets for fine-grained visual classification: FGVC Aircraft, CUB-200-2011, Food-101, Oxford 102 Flowers, Stanford Cars, and Stanford Dogs.

- **FGVC Aircraft (Aircraft)** [16] contains 10,000 images, consisting of 6667 training images and 3333 test images. Each image is annotated with four hierarchical airplane model labels: *model, variant, family, and manufacturer*. We focus on classifying *variant* annotation that includes 100 different subcategories.
- **Stanford Cars (Car)** [15] contains 16,185 images of 196 classes of cars. It is split into 8144 training and 8041 test data. Categories are generally defined according to information about the manufacturer, model, and year of release.
- **Oxford 102 Flowers (Flower)** [39] includes 102 categories of flower images commonly seen in the UK. The training set consists of 20 images per class, and the test set has 6149 images. A single image may contain several flowers. Each image is annotated with a subcategory label.

- **Food-101 (Food)** [40] consists of 101,000 images of 101 food categories. There are manually annotated 250 test images and 750 training images for each class.
- **CUB-200-2011 (CUB)** [14] is the most widely used FGVC dataset. It contains 11,788 images of 200 species of wild birds, which are divided into 5994 for training and 5794 for testing.
- **Stanford Dogs (Dog)** [41] is a collection of 120 different dog categories from around the world. It has 12,000 training data and 8580 test data.

4.2. Implementation Details

By default, the pretrained ViT-S/16 with the DINO framework on ImageNet-1K was used as our initial model. For SSL with FGVC datasets, the models were trained for 5000 iterations, considering the size of the fine-grained dataset, with the AdamW optimizer and a batch size of 256. The learning rate increased linearly during the first 60 iterations to its base value: $lr = 5 \times 10^{-4} \times$ batch_size/256. After this warm-up, the learning rate was decayed using a cosine schedule [42]. Weight decay also followed a cosine schedule from 0.04 to 0.4. We used the same data augmentations as DINO, consisting of random crop, random horizontal flip, random color jittering, Gaussian blur, and solarization. There are two types of views: local and global. For both views, the same data augmentation was used, except for the cropping scale with respect to the original image, to ensure that the local view represents a small region of the image and the global view represents a large region of the image. The projection head h was defined as three-layer MLPs with a 8192 output dimension following DINO. We pretrained and fine-tuned the ViT with 224-sized square images and referenced DINO for most other settings. We used a RoI align layer that outputs a 3×3–sized feature map to compute the consistency loss.

4.3. Experimental Results

Transfer learning. We carried out experiments that froze particular model parameters during transfer learning to investigate how to effectively use the representations pretrained on a large-scale dataset. The comparative results for each dataset are listed in Table 1. The evaluation protocols of k-NN and linear probing were used to examine the quality of the representations.

Table 1. Test accuracy on each dataset under linear probing (k-NN) evaluation. Bold font denotes the best-performing transfer learning method. The "w/o FT" refers to an initial model (i.e., the pretrained model with ImageNet), and the "Full FT" is a model that is fully fine-tuned via SSL on the target FGVC dataset. Note that the transfer learning process is conducted in a self-supervised manner, and the ground-truth labels are only used in the evaluation process.

Transfer Learning (SSL)	**Aircraft [16]**	**Car [15]**	**Flower [39]**	**Food [40]**	**CUB [14]**	**Dog [41]**
w/o FT	61.0 (36.7)	65.8 (22.5)	96.3 (86.5)	79.8 (67.7)	**80.8** (**69.5**)	**83.7** (**77.3**)
Full FT	14.0 (3.5)	50.1 (14.5)	84.5 (72.7)	80.9 (72.7)	56.9 (42.2)	69.0 (62.2)
LayerNorm FT [13]	60.0 (36.4)	66.3 (23.1)	96.9 (88.1)	80.8 (69.8)	76.9 (60.8)	82.1 (74.0)
Lastblock FT	62.3 (37.8)	68.7 (25.8)	**97.2** (90.7)	**82.8** (75.7)	76.1 (60.5)	80.4 (73.6)
Lastblock FT + L_{con}	**63.6** (**39.2**)	**68.9** (**26.8**)	96.9 (**90.9**)	82.6 (**75.9**)	77.1 (60.8)	81.8 (74.2)

As shown in Table 1, "Full FT" demonstrates a degraded performance than "w/o FT" on most of the datasets. Specifically, linear probing performance decreased by an average of 18.7%. However, we observed a slight performance improvement in the Food dataset, which had a relatively sufficient amount of training data. These results confirmed that updating all parameters during SSL with a small-scale dataset might degrade the performance.

However, the overall k-NN and linear probing accuracy are slightly improved when only the layer normalization is fine-tuned using the target dataset. In the Food dataset, there was a performance improvement of 2.1% k-NN accuracy over "w/o FT". Additionally, we examined updating the last block, and the results indicated better performance than fine-tuning only the layer normalization. In particular, the accuracy on Cars increased by 2.4%. These results demonstrate that updating only the last block is more effective when fine-tuned with a small-scale dataset in SSL scenarios. However, the pretrained DINO performed best on the CUB and Dog datasets. We conjecture that because ImageNet contains many dog and bird images, the model can learn adequate representations without further fine-tuning.

We applied the consistency loss term to fine-tune the last block to verify the effectiveness of the proposed consistency loss. The performance can be further enhanced when the consistency loss is combined with the DINO loss, as shown in the last row of Table 1. By considering patch-level consistency, the performance is increased over the absence of the consistency loss in every case except for the linear evaluation on the Flower and Food datasets. In addition, it is possible to mitigate the performance decrease caused by naive fine-tuning on the CUB and Dog datasets. Furthermore, a comparison of the confusion matrices of "LayerNorm FT" and the proposed method on the Aircraft dataset is presented in Figure 3. Positive values in the off-diagonal elements indicate that the misclassification of "LayerNorm FT" is greater than that of the proposed method. Similarly, negative values for the diagonal elements indicate that our proposed method achieves a higher correct classification rate compared with "LayerNorm FT". As shown in Figure 3, the proposed method effectively reduces the number of misclassified samples and improves the correct classification compared with "LayerNorm FT". Specifically, the accuracy of the proposed method is increased in 63 out of 100 classes in the Aircraft dataset. Notably, for classes such as *707-320* and *767-300*, our method achieves 15% higher accuracy compared with "LayerNorm FT".

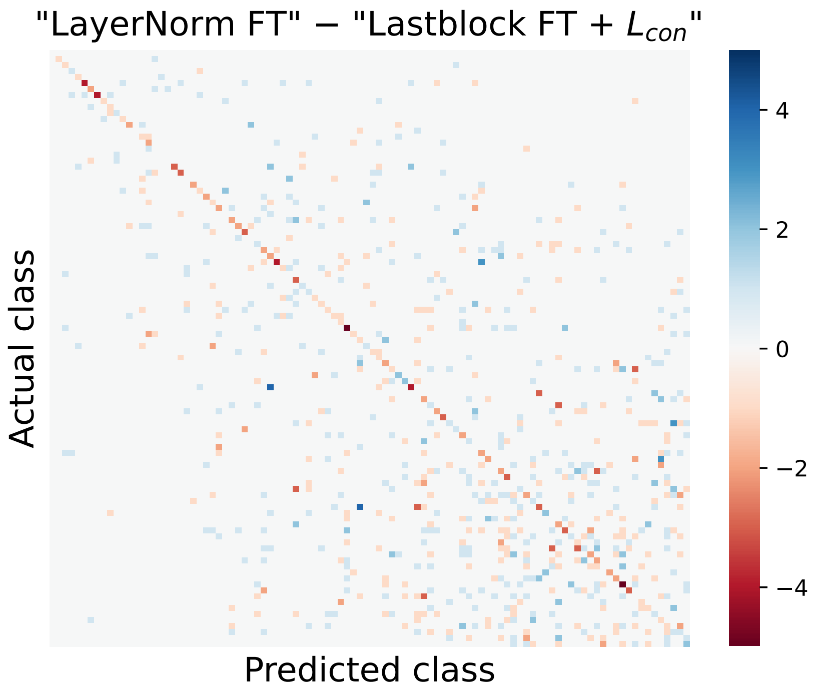


Figure 3. Difference in confusion matrices obtained from linear probed models of "LayerNorm FT" and "Lastblock FT + L_{con}" in Table 1 on Aircraft. Each element is computed by subtracting the corresponding element of the "Lastblock FT + L_{con}" confusion matrix from the "LayerNorm FT" confusion matrix.

We visualized the attention map from fine-tuned models with and without consistency loss to qualitatively demonstrate the effects of consistency loss. The model trained solely with DINO loss assigns high attention scores to both the object and background patches, as shown in Figure 4. On the other hand, objects could be more precisely segmented from the background by applying the consistency loss function.

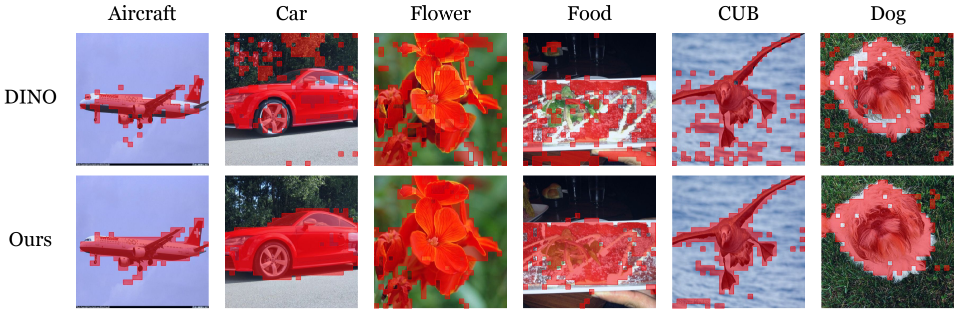


Figure 4. Visualization of masks created by thresholding the self-attention map at 70% of its mass. We demonstrate the resulting masks from fine-tuned models of ViT-S/16 trained with DINO and our method. Specifically, the first row shows the resulting mask for the model with only the last block fine-tuned using the DINO loss. The second row is the visualization result of the model trained with consistency loss under the same setting.

Table 2 contains the fine-grained classification performance of each pretrained model after end-to-end supervised fine-tuning instead of a linear evaluation. At this time, we also consider the model pretrained on ImageNet in a supervised manner, denoted as "Supervised". The SSL pretrained models performed exceptionally better than "Supervised" in the Aircraft, Car, and Flower datasets. In addition, the performance increases when the last block is pretrained during transfer learning to the target dataset (see the third row of Table 2). Specifically, the proposed method shows the best performance among the SSL pretrained models on every FGVC dataset (fourth row of Table 2). On Aircraft, the proposed method improves the performance by 3.96% over "Supervised". In the cases of the Food, CUB, and Dog datasets, the supervised pretrained model performed the best. In summary, the proposed method does not surpass "Supervised" on the Food, CUB, and Dog datasets, but it performs best on all datasets compared with SSL pretrained models. As a result, we confirmed that SSL with the proposed consistency loss helps learn fine-grained representations.

Table 2. Test accuracy on each dataset under fine-tuning evaluation. The first row is the pretrained model with ImageNet in the supervised learning setting. The models in the second and third rows are pretrained using the DINO framework (i.e., the SSL setting). $+L_{con}$ is a pretrained model by applying our proposed consistency loss term.

Pretrained Model	**Aircraft** [16]	**Car** [15]	**Flower** [39]	**Food** [40]	**CUB** [14]	**Dog** [41]
Supervised	81.34	90.15	97.53	**90.97**	**85.36**	**87.74**
Full FT	84.70	91.94	98.10	90.67	82.07	82.33
Lastblock FT	85.05	92.14	98.21	90.78	81.45	82.34
Lastblock FT + L_{con}	**85.30**	**93.10**	**98.54**	90.86	82.29	83.11

ImageNet experiments. The previous experimental results demonstrated the effectiveness of the proposed consistency loss in the transfer learning framework. In this section, we further investigate the effect of consistency loss when applied to a coarse-grained dataset. To this end, we trained ViT-S/16 on the ImageNet dataset from scratch for 300 epochs. Similar to the previous experiments, we set λ and the output size of the RoI align layer to 0.5 and 3×3, respectively.

First, we compared the test accuracy on ImageNet with that of DINO. We pretrained the model on ImageNet with consistency loss and then examined the top 1 accuracy of k-NN classification and linear probing on the validation set of ImageNet. We observed that the proposed consistency loss decreased performance when trained on a large-scale coarse-grained dataset, such as ImageNet. DINO shows 72.8 and 76.1 of test accuracy for k-NN and linear probing, respectively. However, when consistency loss is applied, the accuracy slightly decreases to 71.4 and 74.2 for k-NN and linear probing, respectively. This negative effect is consistent with observations from previous studies [43,44]. Similar to these studies, our method encourages a model to better focus on the small patch regions to capture the subtle patterns. However, for the coarse-grained dataset, these patch regions are more likely to contain irrelevant background information compared with the fine-grained one. Hence, we conjecture that this negative effect might be caused by the use of background information as a shortcut in classification tasks.

Next, the quality of the finer representation of ImageNet pretrained models was evaluated by using the *k*-NN classification accuracy on the FGVC datasets. In Table 3, we compare several SSL methods based on the ViT-S/16 architecture, including MoCov3 [20], iBoT [21], and DINO. The proposed method performed the best among all the SSL methods for the Aircraft, Car, and Flower datasets. Consistent with previous experimental results, the consistency loss was particularly effective on the Aircraft, Car, and Flower datasets. The proposed method demonstrated better performance in half of the fine-grained datasets, although it did not show the highest accuracy across all datasets. When we ranked these methods based on their performance on each dataset, the average rank of the proposed method was the best among the comparison targets.

Table 3. Fine-grained recognition performance of ImageNet pretrained models.

Method	Aircraft [16]	Car [15]	Flower [39]	Food [40]	CUB [14]	Dog [41]	Avg. Rank
MoCo v3 [20]	26.61	16.94	78.45	61.60	44.06	64.71	3.83
iBoT [21]	35.88	20.84	85.02	**68.70**	66.86	**78.39**	2.17
DINO [1]	36.66	22.52	86.45	67.65	**69.45**	64.45	2.17
Ours	**37.32**	**23.22**	**87.54**	66.25	64.36	75.20	1.83

For a qualitative evaluation, we visualized the attention map of the pretrained model with ImageNet. We compared our proposed model with the primary baseline DINO, and the visualization results are shown in Figure 5. From these results, we can observe that our proposed model segments the target object more clearly than DINO. Overall, the model trained with the DINO framework tended to focus on small regions with high attention scores. In contrast, the model trained with consistency loss concentrates on larger areas of an object. These qualitative results demonstrate that the proposed consistency loss helps a model learn image-level representation by aggregating all informative local (i.e., patch-level) representations, not merely focusing on the most discriminative region.

To further investigate the localization capability of the proposed method, we compared the semantic segmentation performance with DINO using the Flower and CUB datasets. Following the previous study [1], the segmentation results are obtained by thresholding the self-attention map from the ImageNet pretrained models, without further training for the segmentation task. As shown in Table 4, our proposed method clearly outperforms DINO on both datasets. These results suggest that the proposed consistency loss enhances the model's focus on object regions (i.e., foreground), revealing its potential applicability to dense prediction tasks, such as object detection and semantic segmentation.

Ablation study. Two hyperparameters are associated with the proposed consistency loss: the weight of the consistency loss λ and the output size of the RoI alignment layer. In this section, we describe an ablation study to further investigate the impact of each hyperparameter. For the RoI pooling size, we examine $\{1 \times 1,\ 3 \times 3\}$ with default consistency loss weight (i.e., $\lambda = 0.5$). For the weight of consistency loss, we consider three values of $\{1, 0.5, 0.1\}$ with a default RoI pooling size of 3×3. Table 5 shows *k*-NN evaluation results for each dataset after fine-tuning using the proposed method. Overall, it is observed that the choice of each hyperparameter does not have a noticeable impact on performance, which confirms that the proposed method is robust to these hyperparameters.

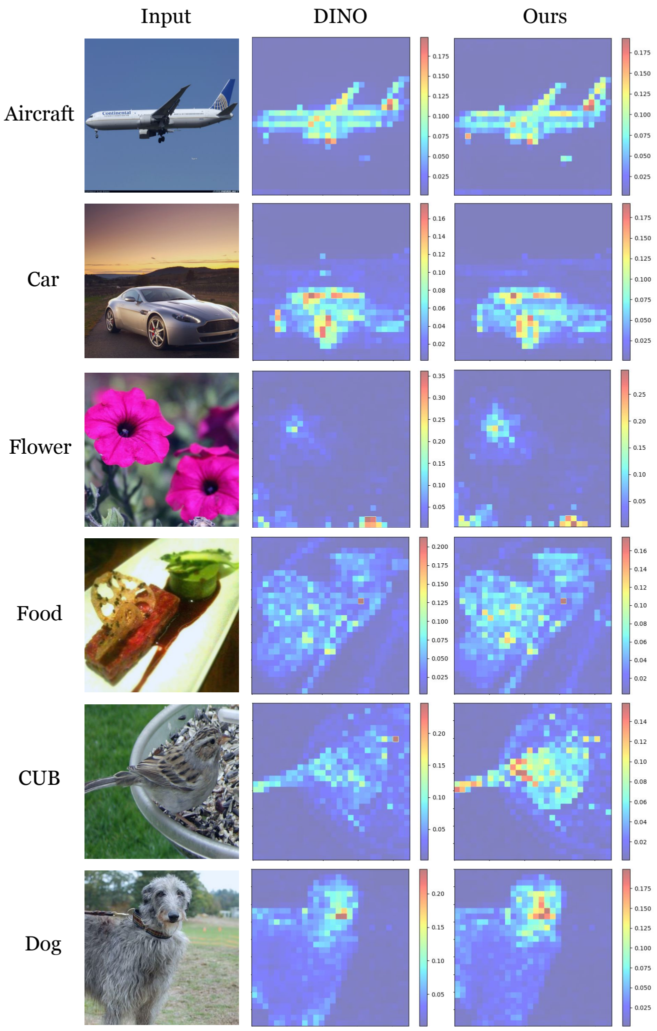


Figure 5. Visualization of attention maps from ImageNet pretrained models. The left column indicates the original images from each FGVC dataset. The middle and the right columns represent the attention maps of DINO and the proposed method, respectively. Six attention maps from the last block are summed and L2-normalized for visualization.

Table 4. Segmentation performance comparison between DINO and our model. We demonstrate results using a pretrained ViT-S/16 with ImageNet and evaluate the Jaccard index between ground truth and predicted masks by applying a 60% self-attention map threshold. The result for the best-performing head is shown.

Method	**Flower** [39]	**CUB** [14]
DINO [1]	29.33	22.02
Ours	**44.87**	**37.64**

Table 5. *k*-NN accuracy of models trained with different λ and pooling size.

λ	**Pooling Size**	**Aircraft** [16]	**Car** [15]	**Flower** [39]	**Food** [40]	**CUB** [14]	**Dog** [41]	**Avg.**
0.5	3 × 3	**39.2**	26.8	**90.9**	**75.9**	**60.8**	74.2	**61.3**
	1 × 1	38.1	**27.6**	89.9	**75.9**	60.6	**74.3**	61.1
1	3 × 3	38.9	26.5	90.5	75.5	**60.9**	**74.6**	61.1
0.5		39.2	26.8	**90.9**	**75.9**	60.8	74.2	**61.3**
0.1		**39.6**	**26.9**	90.5	**75.9**	60.8	74.0	**61.3**

5. Conclusions

In this study, we attempt to apply SSL to small-scale fine-grained data. To this end, we utilize a pretrained ViT with an SSL framework and explore an effective transfer learning strategy. Moreover, we propose consistency loss, which acts as a regularizer to preserve patch-level semantics in the representation space. Extensive experiments on six FGVC benchmark datasets demonstrate that fine-grained representations can be effectively learned using the proposed consistency loss function. Our proposed method can be utilized in various real-world fine-grained image recognition applications that require extensive labeling efforts, including medical image diagnosis, species identification in ecological research, and others.

Although the proposed method offers clear advantages for the FGVC task, it shows lower performance on a coarse-grained dataset such as ImageNet, as reported in previous studies. This issue might be caused by the use of background information as a shortcut in classification tasks. Further investigation is required to determine the causes of these limitations and potential solutions. Meanwhile, direct aggregation of the patch-level representations for the FGVC task could be considered to further improve our method. Exploring effective strategies to integrate the encoded patch information into the final prediction could be an interesting research topic, which we leave for future work. In addition, real-world scenarios often involve the availability of weakly labeled data or limited labeled data. Therefore, extending the proposed method to weakly supervised learning or semisupervised learning settings can be a promising research direction. In this case, further investigation should be conducted to determine effective strategies to fully exploit the limited label information.

Author Contributions: Conceptualization, Y.L. and S.H.; funding acquisition, S.H.; investigation, Y.L., S.L. and S.H.; methodology, Y.L. and S.H.; project administration, S.H.; resources, S.H.; software, Y.L. and S.L.; supervision, S.H.; validation, Y.L. and S.L.; writing—original draft, Y.L.; writing—review and editing, S.L. and S.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT and the Ministry of Education) (NRF-2021R1C1C1011907 and NRF-2019R1A6A1A03032119).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used for this study are publicly available. Refer to the description of each dataset.

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

References

1. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9650–9660.
2. 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.
3. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
4. Liu, Y.; Jin, M.; Pan, S.; Zhou, C.; Zheng, Y.; Xia, F.; Yu, P. Graph self-supervised learning: A survey. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 5879–5900.
5. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
6. Zbontar, J.; Jing, L.; Misra, I.; LeCun, Y.; Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 18–24 July 2021; pp. 12310–12320.
7. Zhang, P.; Wang, F.; Zheng, Y. Self supervised deep representation learning for fine-grained body part recognition. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE, Melbourne, VIC, Australia, 18–21 April 2017; pp. 578–582.

8. Kim, Y.; Ha, J.W. Contrastive Fine-grained Class Clustering via Generative Adversarial Networks. *arXiv* **2021**, arXiv:2112.14971.
9. Wu, D.; Li, S.; Zang, Z.; Wang, K.; Shang, L.; Sun, B.; Li, H.; Li, S.Z. Align yourself: Self-supervised pre-training for fine-grained recognition via saliency alignment. *arXiv* **2021**, arXiv:2106.15788.
10. Cole, E.; Yang, X.; Wilber, K.; Mac Aodha, O.; Belongie, S. When does contrastive visual representation learning work? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14755–14764.
11. Zhao, N.; Wu, Z.; Lau, R.W.; Lin, S. Distilling localization for self-supervised representation learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; pp. 10990–10998.
12. 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 International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
13. Reed, C.J.; Yue, X.; Nrusimha, A.; Ebrahimi, S.; Vijaykumar, V.; Mao, R.; Li, B.; Zhang, S.; Guillory, D.; Metzger, S.; et al. Self-supervised pretraining improves self-supervised pretraining. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2022; pp. 2584–2594.
14. Wah, C.; Branson, S.; Welinder, P.; Perona, P.; Belongie, S. *The Caltech-UCSD Birds-200-2011 Dataset*; California Institute of Technology: Pasadena, CA, USA, 2011.
15. Krause, J.; Stark, M.; Deng, J.; Fei-Fei, L. 3d object representations for fine-grained categorization. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Sydney, Australia, 2–8 December 2013; pp. 554–561.
16. Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv* **2013**, arXiv:1306.5151.
17. Chen, X.; Fan, H.; Girshick, R.; He, K. Improved baselines with momentum contrastive learning. *arXiv* **2020**, arXiv:2003.04297.
18. Oord, A.v.d.; Li, Y.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
19. 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.
20. Chen, X.; Xie, S.; He, K. An empirical study of training self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 9640–9649.
21. Zhou, J.; Wei, C.; Wang, H.; Shen, W.; Xie, C.; Yuille, A.; Kong, T. ibot: Image bert pre-training with online tokenizer. *arXiv* **2021**, arXiv:2111.07832.
22. Chou, P.Y.; Kao, Y.Y.; Lin, C.H. Fine-grained Visual Classification with High-temperature Refinement and Background Suppression. *arXiv* **2023**, arXiv:2303.06442.
23. Lin, D.; Shen, X.; Lu, C.; Jia, J. Deep lac: Deep localization, alignment and classification for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1666–1674.
24. Zhang, H.; Xu, T.; Elhoseiny, M.; Huang, X.; Zhang, S.; Elgammal, A.; Metaxas, D. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1143–1152.
25. Huang, S.; Xu, Z.; Tao, D.; Zhang, Y. Part-stacked CNN for fine-grained visual categorization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1173–1182.
26. Fu, J.; Zheng, H.; Mei, T. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4438–4446.
27. Zhang, T.; Chang, D.; Ma, Z.; Guo, J. Progressive co-attention network for fine-grained visual classification. In Proceedings of the 2021 International Conference on Visual Communications and Image Processing (VCIP), IEEE, Munich, Germany, 5–8 December 2021; pp. 1–5.
28. Zheng, H.; Fu, J.; Mei, T.; Luo, J. Learning multi-attention convolutional neural network for fine-grained image recognition. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5209–5217.
29. He, J.; Chen, J.N.; Liu, S.; Kortylewski, A.; Yang, C.; Bai, Y.; Wang, C. Transfg: A transformer architecture for fine-grained recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2022; Volume 36; pp. 852–860.
30. Sun, H.; He, X.; Peng, Y. Sim-trans: Structure information modeling transformer for fine-grained visual categorization. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portuga, 30 September 2022; pp. 5853–5861.
31. Zhang, Y.; Cao, J.; Zhang, L.; Liu, X.; Wang, Z.; Ling, F.; Chen, W. A free lunch from vit: Adaptive attention multi-scale fusion transformer for fine-grained visual recognition. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Singapore, 22–27 May 2022; pp. 3234–3238.
32. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2020**, *109*, 43–76. [CrossRef]
33. Ribani, R.; Marengoni, M. A survey of transfer learning for convolutional neural networks. In Proceedings of the 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), IEEE, Rio de Janeiro, Brazil, 28–31 October 2019; pp. 47–57.

34. Ayoub, S.; Gulzar, Y.; Reegu, F.A.; Turaev, S. Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning. *Symmetry* **2022**, *14*, 2681. [CrossRef]
35. Malpure, D.; Litake, O.; Ingle, R. Investigating Transfer Learning Capabilities of Vision Transformers and CNNs by Fine-Tuning a Single Trainable Block. *arXiv* **2021**, arXiv:2110.05270.
36. Zhou, H.Y.; Lu, C.; Yang, S.; Yu, Y. ConvNets vs. Transformers: Whose visual representations are more transferable? In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 2230–2238.
37. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 448–456.
38. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
39. 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, IEEE, Bhubaneswar, India, 16–19 December 2008; pp. 722–729.
40. Bossard, L.; Guillaumin, M.; Gool, L.V. Food-101–mining discriminative components with random forests. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 446–461.
41. Khosla, A.; Jayadevaprakash, N.; Yao, B.; Li, F.F. Novel dataset for fine-grained image categorization: Stanford dogs. In Proceedings of the CVPR Workshop on Fine-Grained Visual Categorization (FGVC), Colorado Springs, CO, USA, 25 June 2011; Volume 2, No. 1.
42. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
43. Wang, X.; Zhang, R.; Shen, C.; Kong, T.; Li, L. Dense contrastive learning for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3024–3033.
44. Xiao, T.; Reed, C.J.; Wang, X.; Keutzer, K.; Darrell, T. Region similarity representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 10539–10548.

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.