ConvNets vs. Transformers: Whose Visual Representations are More Transferable?

Hong-Yu Zhou[1]   Chixiang Lu[1]   Sibei Yang[2]   Yizhou Yu[1]

[1]The University of Hong Kong   [2]ShanghaiTech University
{whuzhouhongyu, luchixiang}@gmail.com, yangsb@shanghaitech.edu.cn, yizhouy@acm.org

Abstract

Vision transformers have attracted much attention from computer vision researchers as they are not restricted to the spatial inductive bias of ConvNets. However, although Transformer-based backbones have achieved much progress on ImageNet classification, it is still unclear whether the learned representations are as transferable as or even more transferable than ConvNets' features. To address this point, we systematically investigate the transfer learning ability of ConvNets and vision transformers in 15 single-task and multi-task performance evaluations. Given the strong correlation between the performance of pre-trained models and transfer learning, we include 2 residual ConvNets (i.e., R-101×3 and R-152×4) and 3 Transformer-based visual backbones (i.e., ViT-B, ViT-L and Swin-B), which have close error rates on ImageNet, that indicate similar transfer learning performance on downstream datasets.

We observe consistent advantages of Transformer-based backbones on 13 downstream tasks (out of 15), including but not limited to fine-grained classification, scene recognition (classification, segmentation and depth estimation), open-domain classification, face recognition, etc. More specifically, we find that two ViT models heavily rely on whole network fine-tuning to achieve performance gains while Swin Transformer does not have such a requirement. Moreover, vision transformers behave more robustly in multi-task learning, i.e., bringing more improvements when managing mutually beneficial tasks and reducing performance losses when tackling irrelevant tasks. We hope our discoveries can facilitate the exploration and exploitation of vision transformers in the future.

1. Introduction

Ever since AlexNet [9] was introduced for ImageNet classification [3], convolutional neural networks (i.e., Con-
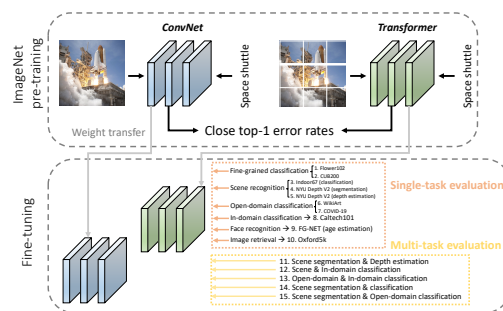

Figure 1: Overview of our investigation procedure. We ask pre-trained ConvNet and Transformer models to have close top-1 error rates on ImageNet classification. The pre-trained weights are then transferred to 15 downstream tasks (i.e., 10 single-task and 5 multi-task duties), to evaluate the transferability of learned representations.

vNets) have become the de-facto choice in computer vision related applications. Over the past decade, researchers have made great efforts to improve the performance of ConvNets, including but not restricted to increasing the network depth with small convolutional kernels [19] and residual connections [6], embedding aggregated multi-branch architectures [20, 26] and automatically searching for neural architectures [28]. Nonetheless, the fundamental constraint of ConvNets, i.e., the inductive bias assumption towards local spatial structures, still remains, making ConvNets naturally disadvantageous in modeling long-range dependencies that are necessary for conducting logical reasoning.

On the other hand, inspired by the attention mechanism [1], Transformers [22] remove convolutional and recurrent operations and solely rely on self-attention to model global dependencies between input and output. Meanwhile, compared to atypical recurrent models, Transformers greatly improve the training efficiency by allowing for large-scale parallelization. Based on above characteristics, Transformers have been the default model choice in various applica-

tions of natural language processing (NLP). In light of the successes that Transformers have achieved, many efforts have been made to extend Transformers to the computer vision field (i.e., building Vision Transformer), where pre-training a general-purpose Transformer-based visual backbone is one of the most promising directions and thus attracts much attention of the community. Vision Transformer (ViT) [4] showed that Transformers can be extended to images to produce competitive ImageNet classification results in comparison to ResNet series [6]. Recently, Liu *et al.* [11] proposed a hierarchical architecture, Swin Transformer, whose representation is computed with shifted windows. Swin Transformer reduces the quadratic computational complexity (with respect to image size) to linear, which promises higher training and inference efficiency. However, although Transformers have achieved results comparable to those of ConvNets on image classification, it is still unclear whether Transformers are able to provide equally transferable representations as ConvNets under the setting of transfer learning.

In this paper, we aim to investigate the transferability of the feature representations of both ConvNets and Transformers on a variety of downstream datasets following a schema of *pre-training first, fine-tuning next*. Note that the scope of fine-tuning could be either the whole network or the last fully-connected layer (refer to linear evaluation protocol in Sec. 3.6). Figure 1 provides an overview of our investigation procedure. Specifically, we first pick two pre-trained ConvNet and Transformer based models, respectively, and require them to have similar top-1 error rates on ImageNet classification. The idea behind is that the accuracy of ImageNet-based pre-training has been shown to have a strong correlation with the accuracy of downstream fine-tuning [8]. Close top-1 errors indicate that the two pre-trained models should presumably have comparable transfer learning performance. Otherwise, if two pre-trained models have quite different performance on ImageNet, the comparison of their fine-tuning performance would be unfair and meaningless as they ought to have some performance differences in transfer learning. Next, we further optimize the pre-trained weights in the fine-tuning stage and evaluate the improved representations on 15 downstream tasks. Different from [27, 17, 17] that evaluate ConvNet's features on single tasks, we conduct both single-task and multi-task learning for more extensive evaluation. The chosen downstream tasks cover a variety of recognition problems, including fine-grained classification (Flower102 and CUB200), indoor scene classification (Indoor67), scene segmentation and depth estimation (NYU Depth V2), in-domain (Caltech101) and open-domain classification (WikiArt and COVID-19).

There are three aspects in our findings. First, transformer-based backbones are more advantageous than

Type	Model	IN (acc.) ↑	Params.	Ave. rank ↓
ConvNet	R-101×3	84.4 (4)	388M	2.5
	R-152×4	85.4 (1)	937M	
Transformer	ViT-B/16	84.0 (5)	86M	3.0
	ViT-L/16	85.2 (2)	307M	
	Swin-B	85.2 (2)	88M	

Table 1: Involved pre-trained models. **/16** denotes the 16×16 input patch size. All models are pre-trained on ImageNet-21k and tested on ImageNet-1k. **IN** is an abbreviation for ImageNet. We display the top-1 accuracy and corresponding performance rank (in a descending order) on Image-1k. For two groups of models (i.e., ConvNet-based and Transformer-based), we present their average ranks, respectively. ↑ denotes the higher the better while ↓ stands for the opposite.

ConvNets when transfer learning is performed on downstream data that have large domain gaps with ImageNet, including but not restricted to fine-grained classification, scene recognition (i.e., classification, segmentation and depth estimation), open-domain classification and face recognition. We believe the above observation provides a strong evidence that Transformer-based backbones produce more generalizable and transferable representations than ConvNet-based models. Second, we observe that the performance advantages (over ConvNets) of two ViT backbones are largely due to whole network fine-tuning, whereas Swin-B does not have such a requirement. Last but not the least, it appears that vision transformers are more robust in multi-task evaluation. More specifically, Transformer-based backbones bring larger improvements when multiple tasks are complementary, while producing smaller performance drops when tasks cannot benefit each other. We believe these advantages can be attributed to two strengths of vision transformers. First of all, they are naturally not confined to the local inductive bias of ConvNets and thus have the ability to capture long-range dependencies. Second, Transformer-based backbones often have much fewer network parameters compared to ConvNets with similar pre-training performance on ImageNet, which would reduce the risk of overfitting when they are transferred to small-scale downstream datasets. When comparing Swin Transformer with two ViT models, we believe the pyramidal feature hierarchy in Swin-B produces more transferable visual features during the pre-training stage and reduces its dependency on whole network fine-tuning.

2. Pre-trained models to be evaluated

There are 5 different pre-trained models that serve as backbones in our experiments, as displayed in Table 1. For ConvNet-based backbones, we choose to use deep residual networks [6] that are among the most effective deep neu-

Model	Flower102 (acc.) ↑	CUB200 (acc.) ↑	Ave. rank ↓
R-101×3	98.4 (5)	87.5 (5)	4.3
R-152×4	98.9 (4)	88.6 (3)	
ViT-B/16	99.2 (3)	88.3 (4)	2.2
ViT-L/16	99.4 (2)	88.9 (2)	
Swin-B	99.8 (1)	89.9 (1)	

Table 2: Results on fine-grained classification. We mark the performance rank (in a descending order) using color gray. ↑ denotes the higher the better while ↓ stands for the opposite.

ral networks with hand-crafted architectures. R-101×3 and R-152×4 are 3× and 4× wider ResNet-101 and ResNet-152, respectively, which are pre-trained with carefully selected strategies [7]. For Transformer-based backbones, we pick ViT-B/16, ViT-L/16 and Swin-B, which are all representative vision transformer backbones. ViT-B/16 and ViT-L/16 comprises alternating layers of multi-head self-attention (MSA) and multi-layer perceptron (MLP) blocks:

$$
\begin{aligned}
\mathbf{z}'_{l+1} &= \text{MSA}(\text{LN}(\mathbf{z}_l)) + \mathbf{z}_l, \\
\mathbf{z}_{l+1} &= \text{MLP}(\text{LN}(\mathbf{z}'_{l+1})) + \mathbf{z}'_{l+1},
\end{aligned}
\tag{1}
$$

where l denotes the layer index and $\mathbf{z}_l$ stands for the input of layer. Swin Transformer improves ViT by replacing layers of MSA and MLP with window-based and shifted-window-based multi-head self-attention (i.e., W-MSA and SW-MSA) with two MLP blocks, which can dramatically reduce the computational complexity:

$$
\begin{aligned}
\tilde{\mathbf{z}}'_{l+1} &= \text{W-MSA}(\text{LN}(\mathbf{z}_l)) + \mathbf{z}_l, \\
\mathbf{z}_{l+1} &= \text{MLP}(\text{LN}(\tilde{\mathbf{z}}'_{l+1})) + \tilde{\mathbf{z}}'_{l+1}, \\
\tilde{\mathbf{z}}'_{l+2} &= \text{SW-MSA}(\text{LN}(\mathbf{z}_{l+1})) + \mathbf{z}_{l+1}, \\
\mathbf{z}_{l+2} &= \text{MLP}(\text{LN}(\tilde{\mathbf{z}}'_{l+2})) + \tilde{\mathbf{z}}'_{l+2}.
\end{aligned}
\tag{2}
$$

From Table 1, we can see that the average performance rank (on ImageNet-1k) of ConvNet-based models is higher than that of Transformer-based. According to the comprehensive investigation in [8], ConvNet-based models would probably achieve better (at least comparable) transfer learning results than Transformer-based networks.

3. Single-task evaluation

We include 10 tasks in single-task evaluation, which consists of a range of topics, such as fine-grained classification, scene recognition, in-domain and open-domain classification, etc. The goal is to extensively evaluate the transferring ability of representations of both ConvNet and Transformer. In the following, we go through different tasks one-by-one. For each experiment, we repeat it for 3 times and report the average results.

3.1. Fine-grained classification

Flower102 [17]. This dataset consists of 102 flower categories that are commonly occurring in UK. Each category contains 40 to 258 images. There are two main challenges of Flower102: i) large similarity between classes and ii) large variation within classes.

CUB200 [23]. 200 bird species and 11,788 images are included in this dataset. The names of species were obtained using an online bird species guide and organized by scientific classification (order, family, genus, species). Flicker Image Search engine is used to acquire bird pictures, which are then filtered by human annotators. There are two versions of CUB200 [24, 23] and we use the latest version [23].

Implementation details. Here we present training and testing strategies of fine-grained classification. Other tasks in this paper may share similar hyper-parameters, where we will clarify the differences of implementation. We conducted all experiments using PyTorch [14].

- Data split: For Flower102, we randomly select 80% images as the training set. The rest 20% data are evenly divided into validation and test sets. For CUB200, we directly use the official test set. 10% images are randomly selected from the official training set to build the validation set, while the rest training images form the training set.

- Network architecture: We append a classification head after each backbone, before which we add a dropout layer (p=0.2).

- Optimizer: Adam is used as the default optimizer, where β_1 is set to 0.9 and β_2 is set to 0.999. We set weight decay to 1e-6.

- Learning rate: The initial learning rate is 1e-4 and is decayed by a factor of 2 each time the validation loss stops decreasing after 3 epochs.

- Augmentation strategies: We use random crop, random rotation (-30 degrees to 30 degrees) and random horizontal flip. The input image size is 224×224.

- Batch size and training time: The training batch size is 32. We stop the training procedure when the validation loss stops decreasing for up to 10 epochs.

- Loss function: We directly use the cross entropy loss.

- Other techniques: We use the label smoothing strategy, where the smoothing rate is set to 0.1. During the fine-tuning stage, we first freeze the whole backbone and conduct warm-up for 200 training iterations using a small learning rate (1e-6). Then, we fine-tune

the whole network using the initial learning rate (i.e., 1e-4).

Results. We present the experimental results in Table 2. Somewhat surprisingly, we discover that Transformer-based backbones hold observable advantages over ConvNets, where the average rank of all 3 vision transformers is much higher than that of 2 ConvNet-based backbones. In contrast with the higher average rank that ConvNets have achieved on ImageNet classification (refer to Table 1), these outstanding results on Flower102 and CUB200 reflect the great discriminative and transferring abilities of transformer-based representations in capturing small differences. More specifically, Swin-B and ViT-L/16 are two best performing backbones on both datasets. ViT-B/16 that achieves the lowest top-1 accuracy on ImageNet-1k produces comparable results with R-152×4 that maintains the highest top-1 accuracy in Table 1. These phenomena again verify the advantages of Transformer when dealing with fine-grained classification problems.

3.2. Scene recognition

Indoor67 [15]. This database contains 67 indoor scene categories, and a total of 15,620 images. The number of images varies across categories, but each category includes at least 100 pictures. All images are collected using online image search engines and some of them come from LabelMe dataset. A minimum resolution of 200 pixels in the smallest axis is guaranteed.

NYU Depth V2 [18]. The dataset comprises rich annotations of both semantic segmentation and depth estimation for 35,064 distinct objects that are collected from 3 different US cities using Kinect. All 1,449 images are divided into 40 scene classes and a dense per-pixel labeling is conducted.

Implementation details. Most details of making experiments on Indoor67 are similar to those of Flower102 and CUB200. In the following, we mainly introduce how to implement indoor semantic segmentation and depth estimation.

- Data split: For Indoor67, we use the official test list while randomly split 10% data from the training list to build the validation set. The rest 90% training data form the training set. For NYU Depth V2 (i.e., both segmentation and depth estimation tasks), we randomly select 80% images as the training set. The validation and test sets include 10% data, respectively.

- Network architecture: For indoor scene segmentation, we directly make use of the segmentation head of UPerNet [25] and append it after ConvNet or Trans-

former backbones. As for depth estimation, we modify the segmentation head to predict depth values.

- Optimizer: We use AdamW [12], where β_1 is set to 0.9, β_2 is set to 0.999 and weight decay is set to 1e-6.

- Learning rate: For both tasks, the initial learning rate is 6e-5 and we employ a polynomial learning rate decay strategy where the power value is set to 0.9.

- Augmentation strategies: For semantic segmentation, we use random crop and random horizontal flip. We also apply a variety of photometric distortion, including but not limited to random brightness, random contrast, random saturation, etc. The input image size is 512×512. For depth estimation, only random crop and random horizontal flip are applied, and the input size 384×384.

- Batch size and training time: The training batch size is 8. The training procedure for each model lasts for 20,000 iterations.

- Loss function: For semantic segmentation, we apply the cross entropy loss (weight=1) and deep supervision loss (weight=0.4). For depth estimation, we employ the scale-and shift-invariant trimmed loss (weight=1) that operates on an inverse depth representation [16] and gradient-matching loss [10] (weight=1).

- Other techniques: We conduct warm-up for 1,500 training iterations using a small learning rate (1e-6). Then, we fine-tune the whole network using the initial learning rate (i.e., 6e-5).

Results. Table 3 displays the experimental results of 3 tasks on scene recognition. Again, Transformer-based backbones prominently outperform ConvNet-based models in the average rank. Moreover, all 3 Transformer-based backbones occupy the top 3 positions on both indoor scene segmentation and depth estimation tasks. Considering recognizing scenes is a quite challenging task that requires strong reasoning ability to understand the relationship between objects and scenes, the obvious improvements brought by vision transformers demonstrate their advantages of mastering complex situations. Besides, we can also observe that Swin-B achieve the best performance on all 3 scene recognition tasks, suggesting the potential of producing high-quality transferable representations using efficient Transformer-based backbones.

3.3. Open-domain classification

WikiArt [21]. This database has a collection of more than 80,000 fine-art paintings from more than 1,000 artists, ranging from the 15-th century to modern times. All images

Model	Indoor67 (acc.) ↑	NYU segmentation (mIoU) ↑	NYU depth estimation (RMSE) ↓	Ave. rank ↓
R-101×3	84.3 (5)	52.2 (5)	0.387 (5)	4.2
R-152×4	85.7 (2)	53.1 (4)	0.382 (4)	
ViT-B/16	85.1 (4)	53.2 (3)	0.368 (3)	2.2
ViT-L/16	85.5 (3)	53.4 (2)	0.360 (2)	
Swin-B	87.6 (1)	54.1 (1)	0.358 (1)	

Table 3: Results on scene recognition. mIoU and RMSE (Root Mean Square Error) are used as the evaluation metrics for segmentation and depth estimation tasks, respectively. We mark the performance rank (in a descending order) using color gray.

Model	WikiArt (acc.) ↑	COVID-19 (acc.) ↑	Ave. rank ↓
R-101×3	66.2 (5)	80.6 (5)	4.5
R-152×4	66.4 (4)	81.2 (4)	
ViT-B/16	67.4 (3)	81.7 (3)	2.0
ViT-L/16	68.4 (2)	82.1 (2)	
Swin-B	71.0 (1)	82.6 (1)	

Table 4: Results on open-domain classification. We mark the performance rank (the smaller the better) using color gray.

Model	Caltech101 (acc.) ↑	Ave. rank
R-101×3	96.8 (2)	2.5
R-152×4	96.7 (3)	
ViT-B/16	96.7 (3)	3.0
ViT-L/16	96.5 (5)	
Swin-B	97.7 (1)	

Table 5: Results on in-domain classification. We mark the performance rank (the smaller the better) using color gray.

are collected from https://www.wikiart.org/ and can be divided into 27 different styles.

COVID-19 Image Data Collection [2]. More than 700 pneumonia cases with chest X-rays are involved, which were built to improve the identification of COVID-19. These X-rays come from over 400 people from 26 countries. In this dataset, we mainly focus on distinguishing COVID-19 from other disease/normal images.

Implementation details. We implement classification networks on WikiArt and COVID-19 Image Data Collection using the same training strategies applied to Flower102 and CUB200. Specifically, in the task of COVID-19 classification, we build the training set using images from Australia and America. X-rays from Africa form the validation set while the remaining images are included in the test set.

Results. The results on open-domain classification are shown in Table 4, from which we can find similar phenomena that have been observed on fine-grained classification and scene recognition. Again and again, Transformer-based backbones surpass ConvNet-based models by large margins. More importantly, we can see that all 3 vision transformers occupy the top 3 places on both WikiArt and COVID-19 as what they have done on tasks of indoor scene segmentation and depth estimation. Considering the images of WikiArt and COVID-19 are not included in ImageNet, the improvements offered by Transformer-based representations provide strong evidences for their great transferability. In addition, we can see that Swin-B again becomes the winner on open-domain classification, verifying its representations are more transferable and generalizable

than those of ViT-L/16 (with the same top-1 accuracy on ImageNet-1k classification).

3.4. In-domain classification

Caltech101 [5]. This dataset consists of 9,146 images, which are acquired by searching names of 101 categories using Google Image Search engine and filtering out irrelevant search results. Each category contains 40 to 800 pictures and most of them have about 50 images. The size of each image is roughly 300 x 200 pixels.

Implementation details. We simply employ the same training strategies on Flower102 and CUB200.

Results. It is not surprising to find ConvNet-based backbones achieve a higher average rank on in-domain classification as some classes and images of Caltech101 overlap with those in ImageNet. If we compare the average ranks in Table 5 with those in Table 1, it is obvious that the they are similar and quite consistent. In other words, we can draw a conclusion that better ImageNet performance often lead to better results on Caltech101 as the backbones have the ability to memorize seen images and recognize similar classes.

3.5. Face recognition

FG-NET[1]**.** FG-NET consists of 1,002 color or gray facial images of more than 50 individuals with large variations in pose, expression and lighting. For each subject, there are more than ten images ranging from age 0 to age 69.

Implementation details. We discuss some specific details about the experiment on FG-Net as follows. For other de-

---

[1] https://yanweifu.github.io/FG_NET_data/

Model	FG-NET (MAE) ↓	Ave. rank ↓
R-101×3	3.6 (3)	
R-152×4	4.7 (5)	4.0
ViT-B/16	3.5 (2)	
ViT-L/16	4.5 (4)	2.3
Swin-B	3.0 (1)	

Table 6: Results on face recognition (i.e., facial age estimation). MAE is defined as the average of the absolute errors between the estimated ages and the ground truth ages. We mark the performance rank (the smaller the better) using color gray.

Model	CUB200 ↑	Indoor67 ↑	WikiArt ↑	Ave. rank ↓
R-101×3	84.7 (2)	83.7 (4)	53.3 (3)	
R-152×4	75.2 (5)	85.6 (2)	53.9 (2)	3.0
ViT-B/16	79.5 (4)	81.8 (5)	47.3 (4)	
ViT-L/16	80.1 (3)	84.1 (3)	46.4 (5)	3.0
Swin-B	87.1 (1)	86.3 (1)	54.2 (1)	

Table 7: Linear classification protocol. The evaluation metric is mean accuracy. We mark the performance rank (the smaller the better) using color gray.

Model	Oxford5k (mAP) ↑	Ave. rank ↓
R-101×3	60.7 (1)	
R-152×4	59.6 (3)	2.0
ViT-B/16	58.2 (4)	
ViT-L/16	57.7 (5)	3.7
Swin-B	59.9 (2)	

Table 8: Results on unsupervised image retrieval. mAP stands for mean average precision. We mark the performance rank (the smaller the better) using color gray.

tails, we simply follow the operations on Flower102 and CUB200.

- Data split: We randomly choose 5 individuals and 3 individuals to build the test and validation sets, respectively. The remaining images are used for training.

- Augmentation strategies: Common augmentation methods like random crop and random horizontal flip are adopted. Besides, we also employ random affine, where the rotation degree is randomly chosen between -10 degrees and 10 degrees and a shear operation (between -12 degrees and 12 degrees) parallel to the x-axis is also applied. The input size is 224×224.

- Loss function: We use two loss functions: mean variance loss [13] (weight=1) and cross entropy loss (weight=1).

Results. Since ImageNet pre-training does not contain many face images, the performance on face recognition can somewhat reflect the transferring and generalization abilities of representations (like the problem of open-domain classification). As shown in Table 6, vision transformers

again achieve a higher average rank than ConvNet-based backbones, demonstrating Transformers are more capable of tackling open-domain problems that are more practical and applicable for real-world applications. Another interesting phenomenon is that bigger models like R-152×4 and ViT-L/16 are outperformed R-101×3 and ViT-B/16, respectively, while Swin-B still takes the first place. Such observations imply that efficient models (with fewer parameters) are more suitable for face recognition tasks as they reduce the risking of overfitting.

3.6. Linear evaluation protocol

In this section, we first perform linear classification using ConvNet- and Transformer-based pre-trained feature representations directly by freezing the backbone and training a supervised linear classification head (i.e., a fully-connected layer followed by softmax) on CUB200, Indoor67 and WikiArt, respectively. In addition, we also conduct unsupervised image retrieval using pre-trained feature representations from each backbone directly.

Implementation details. In linear classification, we train the classifier on the global average pooling features for ConvNet-based backbones and Swin-B. All training details are similar to those of fine-tuning. In unsupervised image retrieval, we first resize each input image to 256×256, after which we apply center crop to generate a central patch whose size is 224×224. Next, we forward each central patch to the backbone network and apply L2 normalization to its extracted feature representation. The evaluation metric is mean average precision (mAP).

Results. Table 7 reports the results of linear classification. Somewhat surprisingly, ConvNet-based backbones achieve a comparable average rank with that of vision transformers. More specifically, it seems that ViT-B/16 and ViT-L/16 lack the ability to provide as transferable representations as they did in fine-tuning. Nonetheless, Swin-B still achieves first places on all 3 datasets, again verifying the effectiveness of feature pyramids. In contrast to the fine-tuning improvements of vision transformers on 3 datasets (presented in Tables 2, 3 and 4), experimental results in Table 7 imply that Transformer-based backbones are more advantageous in fine-tuning than ConvNet-based models. From Table 8, we can see that ConvNet-based backbones have more advantages on the task of unsupervised image retrieval. The underlying reason might be that images in Oxford5 are very similar to those from ImageNet where R-101×3 and R-152×4 produce better classification performance. On the other hand, we notice that Swin-B exhibits much better retrieval results than ViT-B/16 and ViT-L/16. Since Swin-B employs a pyramidal hierarchy scheme to learn visual representations as ConvNets, we think it could be beneficial to

Model	NYU segmentation (mIoU)		NYU depth estimation (RMSE)		Ave. rank ↓
	Perf. ↑	Ave. Imp. ↑ / Drop ↓	Perf. ↓	Ave. Imp. ↑ / Drop ↓	
R-101×3	52.4 (5)		0.381 (5)		
R-152×4	53.5 (4)	0.3	0.379 (4)	0.005	4.5
ViT-B/16	53.9 (3)		0.359 (3)		
ViT-L/16	53.8 (2)	0.4	0.352 (2)	0.007	2.0
Swin-B	54.5 (1)		0.355 (1)		

Table 9: Multi-task learning on scene segmentation and depth estimation. **Perf.** and **Imp.** are abbreviations for performance and improvement, respectively. The green/red color denotes the relative improvement/drop in performance using multi-task learning over using single datasets. The performance ranks (in a descending order) are marked using color grey.

Model	Indoor67 (acc.)		Caltech101 (acc.)		Ave. rank ↓
	Perf. ↑	Ave. Imp. ↑ / Drop ↓	Perf. ↑	Ave. Imp. ↑ / Drop ↓	
R-101×3	82.4 (5)		95.9 (5)		
R-152×4	83.2 (3)	2.2	96.0 (4)	0.8	4.3
ViT-B/16	83.7 (2)		96.9 (3)		
ViT-L/16	82.9 (4)	1.8	97.0 (2)	0.1	3.3
Swin-B	86.3 (1)		97.4 (1)		

Table 10: Multi-task classification on scenes and generic objects. The green/red color denotes the relative improvement/drop in performance using multi-task learning over using single datasets. We mark the performance rank (the smaller the better) using color gray.

learn features for image retrieval in a hierarchical manner using vision transformers.

4. Multi-task evaluation

In this section, we evaluate ConvNet- and Transformer-based backbones on 6 multi-task learning problems. Apart from exact experimental results and corresponding performance ranks, we also present relative performance improvement/drop over single-task learning. Note that we do not strictly require that the individual tasks in each multi-task problem are beneficial to each other (i.e., providing improvements over single tasks). Instead, our goal is to investigate which type of models would better handle the problem of multi-task learning, which is quite necessary and applicable in real-world applications. Similar to single-task evaluation, we repeat each experiment for 3 times and report the average results.

4.1. Scene segmentation and depth estimation

In this setting, each image is associated with segmentation labels and depth values, both of which are acquired from NYU Depth V2 dataset.

Implementation details. We include two heads in UPer-Net, where the original segmentation head is used for scene segmentation and depth prediction head is responsible for depth estimation. For loss functions, we sum up those used in single-task training with equal weights (weight=1). We resize each image to 384×384 and the training batch size is 8. For other details, we directly follow those used in single-task learning.

Results. From Table 9, we can draw a conclusion that scene segmentation and depth estimation are beneficial to each other because the multi-task learning leads to better performance on both tasks over single-task learning. In comparison to ConvNet-based models, vision transformers utilize complementary information hidden in two tasks more effectively, which can be verified by the larger improvements on both tasks (i.e., 0.4 vs. 0.3 on segmentation and 0.007 vs. 0.005 on depth estimation). Not surprisingly, Transformer-based backbones again hold a higher average performance rank than ConvNet-based models.

4.2. Scene and in-domain classification

We perform multi-task classification on a combination of scene and in-domain classification datasets, where each image comes from either Indoor67 or Caltech101, as shown in Table 10.

Implementation details. Images from Indoor67 and Caltech101 are randomly shuffled to build a mixed dataset. During the training stage, we append two classification heads to the backbone, each with a dropout layer (p=0.2). To distinguish images from Indoor67 and those from Caltech101, we add a pre-classification head whose output size is 1. We apply cross entropy loss to train both classification and pre-classification heads. For inference, we take two steps to make predictions. For each test input, we first decide whether it belongs to Indoor67 or Caltech101 according to the output of the pre-classification head (after the sigmoid function). Then, we compute mean accuracy on each dataset, respectively. For other training and inference details, we follow the operations on Flower102 and CUB200.

Model	Caltech101 (acc.)		WikiArt (acc.)		Ave. rank ↓
	Perf. ↑	Ave. Imp. ↑ / Drop ↓	Perf. ↑	Ave. Imp. ↑ / Drop ↓	
R-101×3	90.6 (4)	6.7	66.1 (5)	0.1	4.5
R-152×4	89.6 (5)		66.3 (4)		
ViT-B/16	95.6 (2)		67.6 (3)		
ViT-L/16	92.9 (3)	2.1	67.8 (2)	0.1	2.0
Swin-B	96.1 (1)		71.6 (1)		

Table 11: Multi-task classification on art styles and generic objects. The performance ranks (in a descending order) are marked using color grey, while color green/red denotes the relative improvement/drop over single-task training.

Model	NYU segmentation (mIoU)		Indoor67 (acc.)		Ave. rank ↓
	Perf. ↑	Ave. Imp. ↑ / Drop ↓	Perf. ↑	Ave. Imp. ↑ / Drop ↓	
R-101×3	52.0 (5)	0.3	82.5 (5)	2.1	4.5
R-152×4	52.8 (4)		83.4 (4)		
ViT-B/16	53.2 (3)		83.7 (3)		
ViT-L/16	53.5 (2)	0.1	84.2 (2)	1.3	2.0
Swin-B	53.8 (1)		86.5 (1)		

Table 12: Multi-task learning on scene segmentation and classification. The performance ranks (in a descending order) are marked using color grey. Color green/red denotes the relative improvement/drop over single-task training.

Results. From Table 10, we can see that Transformer-based backbones still maintain observable advantages (i.e., higher performance rank) over ConvNets in the problem of scene and in-domain classification. Specifically, Swin-B again takes the first places on both Indoor67 and Caltech101, demonstrating its capability to deal with scene and in-domain classification, simultaneously. It is not surprising to find that the results on either dataset are slightly lower than those trained on each dataset solely, showing that Indoor67 and Caltech101 are not beneficial to each other. Besides, we find that Transformer-based backbones suffer from smaller performance drops on Indoor67 and perform on-par with single-task classification on Caltech101 using multi-task learning. In comparison, ConvNet-based models suffer from larger performance drops on both datasets.

4.3. Open-domain and in-domain classification

In this part, we replace the scene dataset (i.e., Indoor67) with WikiArt to investigate multi-task learning on a combination of open-domain and in-domain classification problems. Experimental results are presented in Table 11.

Implementation details. We directly refer to the details on scene and in-domain classification (refer to Sec. 4.2).

Results. From Table 11, we can see that Transformer-based backbones outperform ConvNet-based models by large margins in multi-task learning for both art style and generic object recognition. It is understandable that conducting multi-task training using both Caltech101 and WikiArt does not boost the performance over single-task training as the involved images come from two dramatically different domains and thus require different representations towards the goal of recognition. Nonetheless, Transformer-based backbones can still greatly reduce the performance

drop by nearly 5 percents on Caltech101, compared to ConvNets, again verifying the ability of vision transformers in dealing with two non-related tasks.

4.4. Scene segmentation with different classification problems

We investigate the possibility of incorporating scene segmentation into scene classification and open-domain classification, respectively. Experimental results are given in Tables 12 and 13.

Implementation details. For network architecture, we use UPerNet to carry out segmentation. To perform classification tasks at the same time, we replace the fully-connected layers for ImageNet pre-training with classification heads for different classification tasks. Specifically, for ConvNet-based backbones and Swin-B, we add classification heads on global average pooling features in their last layers (i.e., right before the upsampling layer in the bottleneck of UPerNet). The initial learning rate is 1e-4, and we follow the operations in single-task scene segmentation to decrease learning rate and conduct warm-up. The number of training iterations is 20,000. The input size is 384×384. We employ random crop and random horizontal flip as default augmentation strategies.

Results. It is observable that Transformer-based backbones maintain consistent and significant advantages (i.e., higher average performance ranks) over ConvNets in all combinations. Besides, we can still find that it is hard to conduct multi-task learning on top of segmentation and different classification problems even when the incorporated classification problem is closely related to scenes (i.e., Indoor67). Nonetheless, vision transformers are more resistant to tasks that are hard to combine than ConvNets, which

Model	NYU segmentation (mIoU)		WikiArt (acc.)		Ave. rank ↓
	Perf. ↑	Ave. Imp. ↑ / Drop ↓	Perf. ↑	Ave. Imp. ↑ / Drop ↓	
R-101×3	50.9 (3)	3.3	65.7 (5)	0.4	4.3
R-152×4	47.8 (5)		66.1 (4)		
ViT-B/16	51.7 (2)		67.1 (3)		
ViT-L/16	49.2 (4)	2.3	67.3 (2)	0.5	2.2
Swin-B	52.8 (1)		70.8 (1)		

Table 13: Multi-task learning on scene segmentation and open-domain classification. Colors grey, green and red denote the performance ranks, relative performance improvement and drop, respectively.

implies that Transformer may be a better choice than ConvNet when dealing with unknown multi-task problems.

5. Conclusion

We found that Transformer-based backbones provide more transferable representations than ConvNets for fine-tuning, especially when the downstream tasks come from domains very different from ImageNet, which is used for pre-training. Meanwhile, vision transformers are more robust in multi-task learning, where they achieve larger improvements and suffer from smaller performance losses. On the other hand, we observe that ConvNets still have slight advantages on in-domain classification and unsupervised image retrieval. In our future work, we will include more datasets in these two types of problems to obtain more comprehensive results and find out the underlying reasons.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 178–178. IEEE, 2004.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[7] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 491–507. Springer, 2020.

[8] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2661–2671, 2019.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

[10] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018.

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[13] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5285–5294, 2018.

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037, 2019.

[15] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE, 2009.

[16] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *arXiv preprint arXiv:2103.13413*, 2021.

[17] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 806–813, 2014.

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 746–760. Springer, 2012.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[21] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3703–3707. IEEE, 2016.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[23] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[24] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[25] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1492–1500, 2017.

[27] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27:3320–3328, 2014.

[28] Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.