Revealing the Dark Secrets of Masked Image Modeling

Zhenda Xie*[13], Zigang Geng*[23], Jingcheng Hu[13], Zheng Zhang[3], Han Hu[3], Yue Cao[3†]
[1]Tsinghua University [2]University of Science and Technology of China [3]Microsoft Research Asia
{t-zhxie, t-ziganggeng, v-jingchu, zhez, hanhu, yuecao}@microsoft.com

Abstract

Masked image modeling (MIM) as pre-training is shown to be effective for numerous vision downstream tasks, but how and where MIM works remain unclear. In this paper, we compare MIM with the long-dominant supervised pre-trained models from two perspectives, the visualizations and the experiments, to uncover their key representational differences. From the visualizations, we find that MIM brings locality inductive bias to all layers of the trained models, but supervised models tend to focus locally at lower layers but more globally at higher layers. That may be the reason why MIM helps Vision Transformers that have a very large receptive field to optimize. Using MIM, the model can maintain a large diversity on attention heads in all layers. But for supervised models, the diversity on attention heads almost disappears from the last three layers and less diversity harms the fine-tuning performance. From the experiments, we find that MIM models can perform significantly better on geometric and motion tasks with weak semantics or fine-grained classification tasks, than their supervised counterparts. Without bells and whistles, a standard MIM pre-trained SwinV2-L could achieve state-of-the-art performance on pose estimation (78.9 AP on COCO test-dev and 78.0 AP on CrowdPose), depth estimation (0.287 RMSE on NYUv2 and 1.966 RMSE on KITTI), and video object tracking (70.7 SUC on LaSOT). For the semantic understanding datasets where the categories are sufficiently covered by the supervised pre-training, MIM models can still achieve highly competitive transfer performance. With a deeper understanding of MIM, we hope that our work can inspire new and solid research in this direction. Code will be available at https://github.com/zdaxie/MIM-DarkSecrets.

1. Introduction

Pre-training of effective and general representations applicable to a wide range of tasks in a domain is the key to the success of deep learning. In computer vision, supervised classification on ImageNet [14] has long been the dominant pre-training task which is manifested to be effective on a wide range of vision tasks, especially on the semantic understanding tasks, such as image classification [17, 18, 36, 38, 51], object detection [23, 29, 61, 63], semantic segmentation [53, 69], video action recognition [7, 52, 65, 67] and so on. Over the past several years, "masked signal modeling", which masks a portion of input signals and tries to predict these masked signals, serves as a universal and effective self-supervised pre-training task for various domains, including language, vision, and speech. After (masked) language modeling repainted the NLP field [15, 49], recently, such task has also been shown to be a competitive challenger to the supervised pre-training in computer vision [3, 8, 18, 27, 74, 80]. That is, masked image modeling (MIM) pre-trained models achieve very high fine-tuning accuracy on a wide range of vision tasks of different nature and complexity.

However, there still remain several questions:

1. What are the key mechanisms that contribute to the excellent performance of MIM?

2. How transferable are MIM and supervised models across different types of tasks, such as semantic understanding, geometric and motion tasks?

To investigate these questions, we compare MIM with supervised models from two perspectives, the visualization perspective and the experimental perspective, trying to uncover key representational differences between these two pre-training tasks and deeper understand the behaviors of MIM pre-training.

We start with studying the attention maps of the pre-trained models. Firstly, we visualize the averaged attention distance in MIM models, and we find that **masked image modeling brings locality inductive bias to the trained model, that the models tend to aggregate near pixels in part of the attention heads,** and the locality strength is highly correlated with the masking ratio and masked patch size in the pre-training stage. But the supervised models tend to focus locally at lower layers but more globally at higher layers.

[*] Equal Contribution. The work is done when Zhenda Xie, Zigang Geng, and Jingcheng Hu are interns at Microsoft Research Asia. [†] Contact person.

We next probe how differently the attention heads in MIM trained Transformer behave. We find that **different attention heads tend to aggregate different tokens on all layers in MIM models**, according to the large KL-divergence on attention maps of different heads. But for supervised models, the diversity on attention heads diminishes as the layer goes deeper and almost disappears in the last three layers. We drop the last several layers for supervised pre-trained models during fine-tuning and find that it benefits the fine-tuning performance on downstream tasks, however this phenomenon is not observed for MIM models. That is, **less diversity on attention heads would somewhat harm the performance on downstream tasks**.

Then we examine the representation structures in the deep networks of MIM and supervised models via the similarity metric of Centered Kernel Alignment (CKA) [37]. We surprisingly find that **in MIM models, the feature representations of different layers are of high similarity, that their CKA values are all very large (e.g., [0.9, 1.0]).** But for supervised models, as in [59], different layers learn different representation structures, that their CKA similarities vary greatly (e.g., [0.5,1.0]). To further verify this, we load the pre-trained weights of randomly shuffled layers during fine-tuning and find that supervised pre-trained models suffer more than the MIM models.

From the experimental perspective, a fundamental pre-training task should be able to benefit a wide range of tasks, or at least it is important to know for which types of tasks MIM models work better than the supervised counterparts. To this end, we conduct a large-scale study by comparing the fine-tuning performance of MIM and supervised pre-trained models, on three types of tasks, semantic understanding tasks, geometric and motion tasks, and the combined tasks which simultaneously perform both.

For semantic understanding tasks, we select several representative and diverse image classification benchmarks, including Concept Generalization (CoG) benchmark [62], the widely-used 12-dataset benchmark [38], as well as a fine-grained classification dataset iNaturalist-18 [68]. For the classification datasets whose categories are sufficiently covered by ImageNet categories (e.g. CIFAR-10/100), supervised models can achieve better performance than MIM models. However, for other datasets, such as fine-grained classification datasets (e.g., Food, Birdsnap, iNaturalist), or datasets with different output categories (e.g., CoG), most of the representation power in supervised models is difficult to transfer, thus MIM models remarkably outperform supervised counterparts.

For geometric and motion tasks that require weaker semantics and high-resolution object localization capabilities, such as pose estimation on COCO [48] and CrowdPose [44], depth estimation on NYUv2 [64] and KITTI [22], and video object tracking on GOT10k [32], TrackingNet [55], and LaSOT [20], MIM models outperform supervised counterparts by large margins. Note that, without bells and whistles, Swin-L with MIM pre-training could achieve state-of-the-art performance on these benchmarks, e.g., 80.5 AP on COCO *val*, 78.9 AP on COCO *test-dev*, and 78.0 AP on CrowdPose of pose estimation, 0.287 RMSE on NYUv2 and 1.966 RMSE on KITTI of depth estimation, and 70.7 SUC on LaSOT of video object tracking.

We select object detection on COCO as the combined task which simultaneously performs both semantic understanding and geometric learning. For object detection on COCO, MIM models would outperform supervised counterparts. Via investigating the training losses of object classification and localization, we find that MIM models help localization task converge faster, and supervised models benefit more for object classification, that categories of COCO are fully covered by ImageNet.

In general, MIM models tend to exhibit improved performance on geometric/motion tasks with weak semantics or fine-grained classification tasks compared to their supervised counterparts. For tasks/datasets where supervised models excel in transfer, MIM models can still achieve competitive transfer performance. Masked image modeling appears to be a promising candidate for a general-purpose pre-trained model. We hope our paper contributes to this understanding within the community and stimulates further research in this direction.

2. Background

Masked Image Modeling. Masked image modeling (MIM) is a sub-task of masked signal prediction, that masks a portion of input images, and lets the deep networks predict the masked signals conditioned on the visible ones. We use SimMIM [74], a simple framework for masked image modeling, as the exampled framework of pre-trained image models in our visualizations and experiments, because it is simple, effective, and generally applicable. Note that, the SimMIM framework could be directly applied to different types of backbone architectures, such as Vision Transformer (ViT) [18], Swin Transformer [51], and ConvNets [16, 30]. This property enables us to study the characteristics of MIM under different types of backbone architectures, as well as in multiple types of downstream tasks.

Other Pre-training Methods. Supervised pre-training [16, 30, 51, 66] is the primary method we employ for comparison with MIM. Contrastive learning approaches [6, 9, 10], as the most successful self-supervised learning methods prior to MIM, have demonstrated exceptional transfer performance and generalization capabilities. Pixel-level pretext tasks [70, 72, 73], on the other hand, exhibit outstanding performance in dense prediction downstream tasks. Since the advent of MIM, several works have attempted to use masking as a data augmentation strategy [1] or to combine MIM with contrastive learning [80], resulting in additional gains. *We provide a more comprehensive analysis of various methods through visualizations and experimental evaluations in Appendix.*

Backbone Architectures. Masked image modeling is mostly studied in the Transformer architectures, thus the major understandings and experiments in this paper are performed on Vision Transformers (ViT) [18] and Swin Transformers [50, 51]. Due to the simple and clear architecture designs of ViT, most of the visualizations in the main paper are performed on ViT, shown in Section 3. Due to the general-purpose property of Swin Transformer, most of the experiments on different downstream tasks in the main paper are conducted on Swin Transformer, shown in Section 4. We provide more visualizations and experimental results in Appendix.

3. Visualizations

3.1. Revealing the Properties of Attention Maps

Attention mechanism [2] has been an exceptional component in deep networks. It is naturally interpretable since attention weights have a clear meaning: how much each token is weighted when determining the output representation of the current token. Fortunately, most MIM pre-trained models [3, 18, 27, 74, 80] are established upon the Vision Transformers, where self-attention block is its major component. Here we start with studying the attention maps of the pre-trained models from three angles: (a) averaged attention distance to measure whether it is local attention or global attention; (b) entropy of attention distribution to measure whether it is focused attention or broad attention; (c) KL divergence of different attention heads to investigate that attention heads are attending different tokens or similar ones.

3.1.1 Local Attention or Global Attention?

Images are observed to exhibit strong locality: pixels near each other tend to be highly correlated [33], motivating the use of local priors in a wide range of visual perception architectures [21,30,40,42,51]. In the era of Vision Transformers, the usefulness of local priors has still undergone rich discussions and trials [18,45,51]. Thus it is valuable to investigate whether MIM models bring the locality inductive bias to the models. We do this by computing averaged attention distance in each attention head of each layer.

Results of the averaged attention distance in different attention heads (dots) w.r.t the layer number, on supervised model (DeiT), contrastive learning model (MoCo v3) and SimMIM model with ViT-B as backbone are shown in Figure 1. We find that the supervised model tends to focus locally at lower layers but more globally at higher layers, which well matches the observations in ViT [18]. Surprisingly, the contrastive learning model acts very similarly to the supervised counterpart. And this similarity may lead to high linear evaluation accuracy on ImageNet-1K of MoCo v3 (76.7% of top-1 accuracy). But for the model trained by SimMIM, its behavior is significantly different to supervised and contrastive learning models. Each layer has diverse attention heads that tend to aggregate both local and global pixels, and the average attention distance is similar to the lower layers of the supervised model. As the number of layers gets deeper, the averaged attention distance becomes even slightly smaller. That is, MIM brings locality inductive bias to the trained model, that the models tend to aggregate near pixels in part of the attention heads. Also, a similar observation could be observed with Swin-B as the backbone, as shown in Figure 2(b).

SimMIM [74] designed a new metric, AvgDist, which measures the averaged Euclidean distance of masked pixels to the nearest visible ones and indicates the task difficulty and effectiveness of MIM depending on the masking ratio and masked patch size. As shown in Figure 2(a), AvgDist is a good indicator that the entries of high fine-tuning accuracy roughly distribute in a range of [10, 20] of AvgDist, while entries with smaller or higher AvgDist perform worse. Interestingly, in the range of [10, 20] of AvgDist, we can also observe a small averaged attention distance. That is, a moderate prediction distance in MIM will bring a greater strength of locality and incur a better fine-tuning performance.

3.1.2 Focused Attention or Broad Attention?

We then measure the attention maps on whether attention heads focus on a few tokens or attend broadly over many tokens, via averaging the entropy of each head's attention distribution. Results of entropy values w.r.t different layers of three pre-trained models, supervised model (DeiT), contrastive learning model (MoCo v3), and MIM model (SimMIM) with ViT-B as the backbone, are shown in Figure 3. For supervised models, we find that some attention heads in lower layers have very focused attention, but in higher layers, most attention heads focus very broadly. The contrastive model still behaves very similarly to the supervised model. But for the MIM model, the entropy values in different attention heads are diverse in all layers, that some attention heads are more focused and some heads have very broad attention.

3.1.3 Diversity on Attention Heads

From the previous two sub-sections, we observe a similar phenomenon, that is, for the supervised model, the attention distance or entropy of attention heads in the last few layers seem to be similar, while for the MIM model, different heads in all layers behave more diversely. Therefore, we want to further explore whether the different heads pay attention to

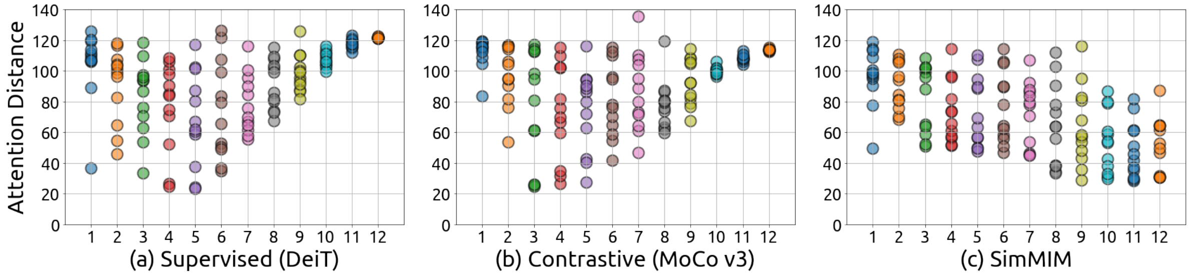

Figure 1. The averaged attention distance in different attention heads (dots) w.r.t the layer number on supervised model (a), contrastive learning model (b), and SimMIM model (c) with ViT-B as the backbone architecture.

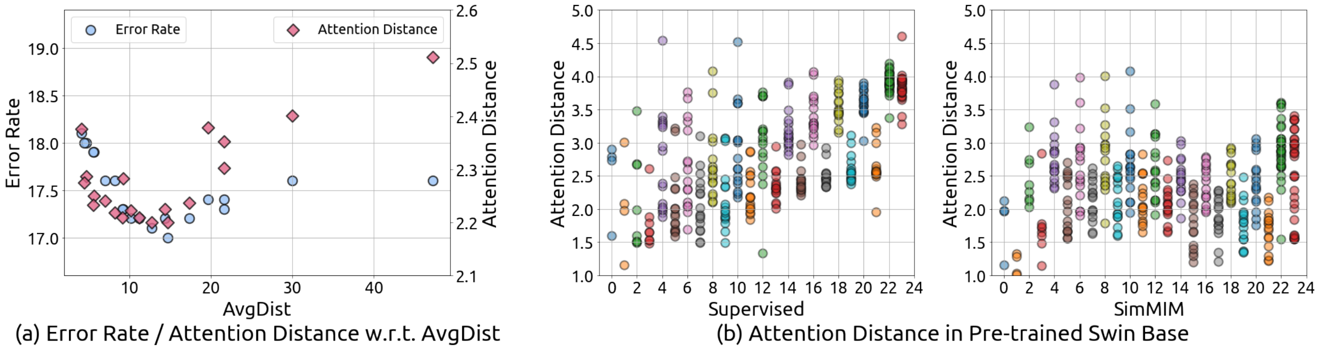

Figure 2. (a) The error rate of fine-tuning on ImageNet-1K (blue circle ○) and averaged attention distance (red diamond ◇) w.r.t AvgDist (averaged distance of masked pixels to the nearest visible pixels) with Swin-B as the backbone. Points (◇ or ○) denote the SimMIM models with different masking ratios and masked patch sizes. (b) The averaged attention distance in different attention heads (dots) w.r.t the layer number on supervised model (b1) and SimMIM model (b2) with Swin-B as the backbone.

different/similar tokens, via computing the Kullback–Leibler (KL) divergence [41] between the attention maps of different heads in each layer.

Results of KL divergence between attention distributions of different heads w.r.t different layers of three pre-trained models, supervised model (DeiT), contrastive learning model (MoCo v3), and MIM model (SimMIM) with ViT-B as the backbone, are shown in Figure 4. As we expect, different attention heads tend to aggregate different tokens on all layers in MIM models, according to the large KL-divergence on attention maps of different heads. But for supervised models and contrastive learning models, the diversity on attention heads becomes smaller as the layer goes deeper and almost disappears from the last three layers.

Intuitively, losing diversity across different attention heads may limit the capacity of the model. To investigate whether the loss of diversity on attention heads has any adverse effect, we gradually drop layers from the end, and only load previous layers when fine-tuning the model for the downstream tasks of COCO $val2017$ pose estimation and NYUv2 depth estimation. From Figure 5, we can observe that when we drop two to eight layers, although the model becomes smaller, the performance of the supervised pre-trained model on COCO $val2017$ pose estimation is better than the baseline, and the performance on NYUv2 depth estimation is comparable with the baseline. This shows that in the supervised pre-trained model, the last layers with small diversity on attention heads indeed affect the performance of downstream tasks. The detailed setup of this experiment is in the Appendix.

3.2. Investigating the Representation Structures via CKA similarity

Studying the behaviors of attention mechanisms is analyzing inside the block, from a micro perspective. Next, we hope to study from a macro perspective of deep networks, such as studying the similarity between feature maps across different layers via the CKA similarity [37]. Results of CKA similarity between feature representations of different layers of three pre-trained models, supervised model (DeiT), contrastive learning model (MoCo v3), and MIM model (SimMIM) with ViT-B as the backbone, are shown in Figure 6. We surprisingly find that in MIM models, the representation structures of different layers are almost the same, that their CKA similarities are all very large (e.g., [0.9, 1.0]). But for supervised models, as in [59], different layers learn different representation structures, that their CKA similarities vary greatly (e.g., [0.5,1.0]). Different from previous

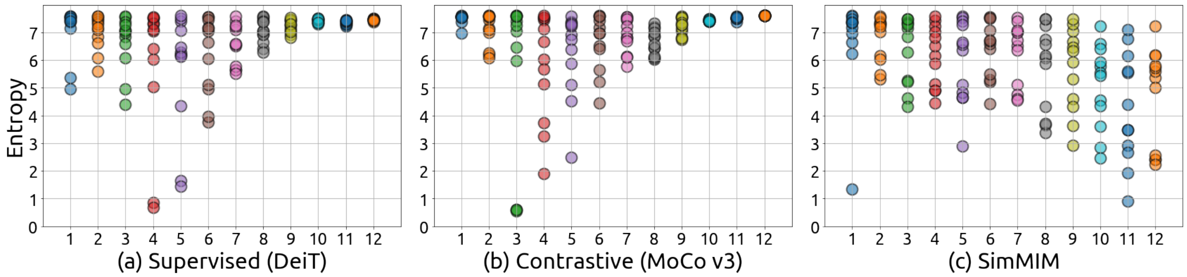

Figure 3. The entropy of each head's attention distribution w.r.t the layer number on (a) supervised model, (b) contrastive learning model, and (c) SimMIM model with ViT-B as the backbone.

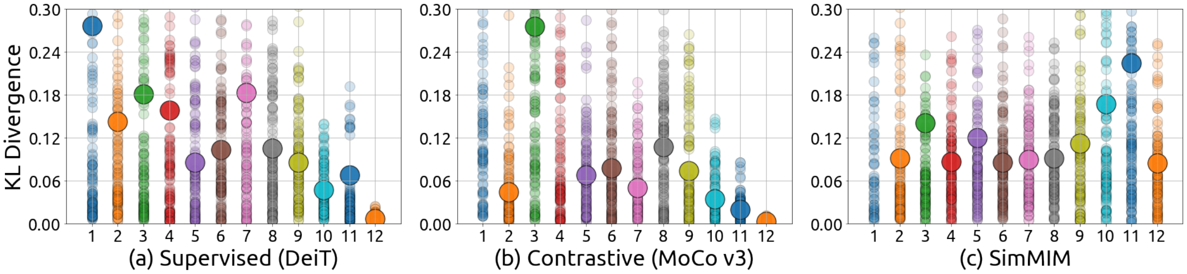

Figure 4. The KL divergence between attention distributions of different heads (small dots) and the averaged KL divergence (large dots) in each layer w.r.t the layer number on (a) supervised model, (b) contrastive learning model, and (c) SimMIM model with ViT-B as the backbone architecture.

visualizations, MoCo v3 behaves similarly to SimMIM in this case.

To further verify this observation, we load the pre-trained weights of randomly shuffled layers and fine-tune the model for the downstream tasks of COCO pose estimation and NYUv2 depth estimation. We observe that by loading the models with the randomly sampled layers, the performance on 1K-MIM drops from 75.5 to 75.2 (-0.3) on pose estimation and 0.382 to 0.434 (-0.052) on depth estimation. But supervised pre-trained models suffer more than the MIM models, which drops from 75.8 to 74.9 (-0.9) on pose estimation, and 0.376 to 0.443 (-0.067) on depth estimation. The detailed setup of this experiment is in the Appendix.

4. Experimental Analysis on Three Types of Downstream Tasks

In this section, we conduct a large-scale study by comparing the fine-tuning performance of MIM and supervised pre-trained models, on three types of tasks, semantic understanding tasks (e.g., image classification in different domains), geometric and motion tasks (e.g., pose/depth estimation, and video object tracking), and the combined tasks which simultaneously perform both types of tasks (e.g., object detection). We use 8 NVIDIA V100 GPUs for our experiments.

4.1. Semantic Understanding Tasks

For semantic understanding tasks, we select several representative and diverse image classification benchmarks, including Concept Generalization (CoG) benchmark [62], the widely-used 12-dataset benchmark [38], as well as a fine-grained classification dataset iNaturalist-18 [68].

Setup. The CoG benchmark consists of five 1k-category datasets split from ImageNet-21K, which has an increasing semantic gap with ImageNet-1K, from L_1 to L_5. On the CoG dataset, we search for the best hyper-parameters based on the top-1 accuracy of the L_1 validation set and then apply the best setting to CoG L_2 to L_5 to report the top-1 accuracy. On the K12 dataset, we adopt standard splits of train/val/test sets as in [38]. We use the training set to fine-tune the models, use the validation set to search for the best hyper-parameters, and then train the models on the merged training and validation sets using the best setting. Following [38], we report mean-per-class accuracy for Aircraft, Pets, Caltech-101, Oxford 102 Flowers and top-1 accuracy for other datasets. The iNat18 dataset includes 437,513 training images and 24,426 validation images, with more than 8,000 categories. We fine-tune the pre-trained models using the training set and report the top-1 accuracy on the validation set. For all datasets, we choose learning rate, weight decay, layer decay, and DropPath [31] on the valid set respectively for the MIM pre-trained model and the supervised pre-trained model. We

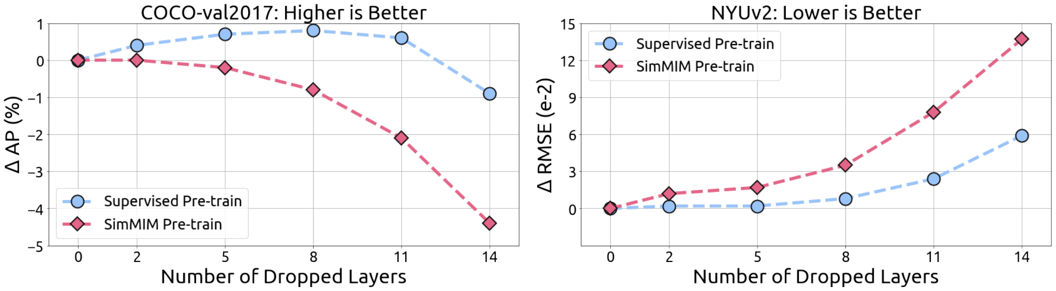

Figure 5. The performance of the COCO $val2017$ pose estimation (left) and NYUv2 depth estimation (right) when we drop several last layers of the SwinV2-B backbone. When the model becomes smaller, the performance of the supervised pre-trained model increases on the pose estimation and keeps the same on the depth estimation. The last layers in the supervised pre-trained model lose diversity across different attention heads and are harmful to the downstream tasks.

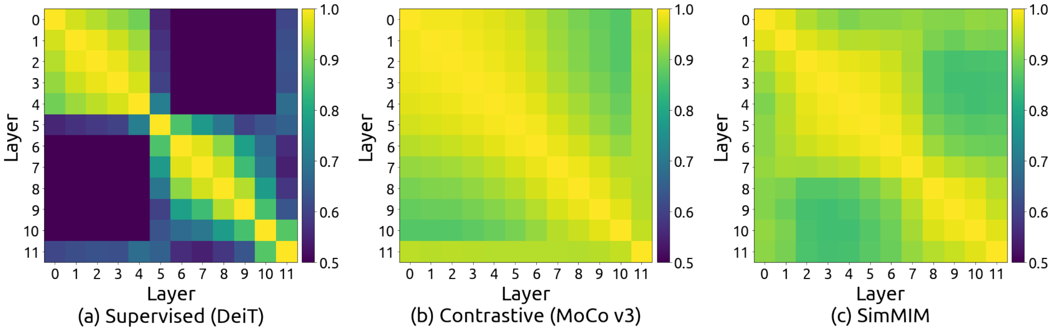

Figure 6. The CKA heatmap between the feature maps of different layers of (a) supervised model, (b) contrastive learning model, and (c) SimMIM model with ViT-B as the backbone architecture.

use the AdamW optimizer [54] and cosine learning rate schedule. We train the model for 100 epochs with 20 warm-up epochs. We adopt the AutoAug [12] and Mixup [78] data augmentation and the input image size is 224 × 224. Other detailed setups of these datasets are in the Appendix.

Results. Results of different semantic understanding tasks are shown in Table 1. For the classification datasets whose categories are sufficiently covered by ImageNet categories (e.g. CIFAR-10/100), supervised models can achieve better performance than MIM models as pre-training. However, for other datasets, such as fine-grained classification datasets (e.g., Food, Birdsnap, iNaturalist), or datasets with different output categories (e.g., CoG), most of the representation power in supervised models is difficult to transfer; thus MIM models remarkably outperform supervised counterparts.

4.2. Geometric and Motion Tasks

We study how MIM models perform on the geometric and motion tasks that require the ability to localize the objects and are less dependent on semantic information. We select several benchmarks, such as pose estimation on COCO [48] and CrowdPose [44], depth estimation on NYUv2 [64] and KITTI [22], and video object tracking on GOT10k [32], TrackingNet [55], and LaSOT [20].

Setup. For pose estimation on COCO and Crowdpose, we use the standard splits for training and evaluation and report the AP based on OKS as the evaluation metric. We use the standard person detection results from [71]. We follow Simple Baseline [71], which upsamples the last feature of the backbone by deconvolutions and predicts the heatmaps at 4× resolution. The data augmentations include random flipping, half body transformation, random scale, random rotation, grid dropout, and color jittering. The input image size is 256 × 256 by default. We use the AdamW [54] optimizer with the base learning rate $5e$-4 and the weight decay $5e$-2. The learning rate is dropped to $5e$-5 at the $120th$ epoch. We train the models for 150 epochs. We use a layer decay of 0.9/0.85 for Swin-B/L and the DropPath [31] of 0.3/0.5 for Swin-B/L.

For depth estimation on NYUv2 and KITTI, we use the standard splits and report the RMSE (Root Mean Square Error) as the evaluation metric. To compare with the previous works [34, 60], we set the maximum range as 10m/80m for NYUv2/KITTI. The head of the depth estimation is the same as that of the pose estimation and is comprised of deconvolutions. Similar to the GLPDepth [34], we use the fol-

pre-train	Concept Generalization (CoG)					Kornlith et al's 12 datasets (K12)					iNat18
	L_1	L_2	L_3	L_4	L_5	Food	Birdsnap	Cars	Aircraft	Average (7)	
1K-SUP	79.4	76.2	72.7	72.5	68.4	93.2	81.8	88.6	83.0	89.7	77.7
1K-MIM	79.6	77.1	73.6	73.0	69.1	94.2	83.7	89.2	83.5	86.1	79.6

Table 1. Comparisons of MIM and supervised (SUP) pre-trained models on semantic understanding tasks with SwinV2-B as the backbone. We follow [38] to report top-1 accuracy (↑) and mean per-class accuracy (↑) for specific datasets. Results on the multi-label dataset Pascal Voc 2007 are not included, whose evaluation metric is not compatible with others.

lowing data augmentations: random horizontal flip, random brightness/gamma/hue/saturation/value and random vertical CutDepth. We randomly crop the images to 480 × 480 / 352 × 352 size for NYUv2/KITTI dataset. The optimizer, layer decay, and DropPath is the same as the pose estimation. The learning rate is scheduled via polynomial strategy with a factor of 0.9 with a minimal value of $3e$-5 and a maximum value of $5e$-4. The total number of epochs is 25. We use the flip testing and sliding window test.

Following the previous methods [13, 47], we train the models on the train splits of four datasets GOT10k [32], TrackingNet [55], LaSOT [20], and COCO [48] and report the success score (SUC) for the TrackingNet dataset and LaSOT dataset, and the average overlap (AO) for GOT10k. We use the SwinTrack [47] to train and evaluate our pre-trained models with the same data augmentations, training, and inference settings. We sample 131072 pairs per epoch and train the models for 300 epochs. We use the AdamW optimizer with a learning rate of $5e$-4 for the head, a learning rate of $5e$-5 for the backbone, and a weight decay of $1e$-4. We decrease the learning rate by a ratio of 0.1 at the 210th epoch. We set the sizes of search images and templates as 224 × 224 and 112 × 112.

Results. From Table 2, for the pose estimation, MIM models pre-trained with ImageNet-1K surpass supervised counterparts by large margins, 2.4 AP on COCO val, 2.2 AP on COCO $test\text{-}dev$, and 4.2 AP on CrowdPose dataset which contains more crowded scenes. Even if the supervised models are pre-trained with ImageNet-22K, the performances are still worse than MIM models pre-trained with ImageNet-1K. The observation of the SwinV2-L is similar to that of the SwinV2-B. With a larger image size 384 × 384, MIM pre-trained SwinV2-L reaches 78.4 on COCO $test\text{-}dev$, and 77.1 on the challenging CrowdPose dataset. Using a stronger detection result from BigDetection [4], we obtain 80.5 AP on COCO val, 78.9 AP on COCO $test\text{-}dev$, and 78.0 AP on CrowdPose.

For the depth estimation, using a simple deconvolution head, SwinV2-B with MIM pre-training with ImageNet-1K achieves 0.304 RMSE on NYUv2 and 2.050 RMSE on KITTI, outperforming the previous SOTA method BinsFormer-L [46]. The MIM pre-training does improve the performance of SwinV2-B by 0.03 RMSE compared with the supervised pre-training with ImageNet-22K. Note that with supervised pre-training, a larger model SwinV2-L shows no gain for the NYUv2 dataset, while with MIM pre-training, SwinV2-L leads to about 0.02 RMSE gain over SwinV2-B.

For the video object tracking, MIM models also show a stronger transfer ability over supervised pre-trained models. On the long-term dataset LaSOT, SwinTrack [47] with MIM pre-trained SwinV2-B backbone achieves comparable result with the SOTA MixFormer-L [13] with a larger image size 320 × 320. We obtain the best SUC of 70.7 on the LaSOT with SwinV2-L backbone with the input image size 224 × 224 and template size 112 × 112.

4.3. Combined Task of Object Detection

We select object detection on COCO as the combined task which simultaneously performs both semantic understanding and geometric learning. For object detection on COCO, a Mask-RCNN [29] framework is adopted and trained with a 3× schedule (36 epochs). We utilize an AdamW [35] optimizer with a learning rate of 6e-5, a weight decay of 0.05, and a batch size of 32. We employ a large jittering augmentation (1024 × 1024 resolution, scale range [0.1, 2.0]).

On COCO, we could clearly observe that MIM model outperforms its supervised counterpart (52.9/46.7 v.s. 51.9/45.7 of box/mask AP) with SwinV2-B as the backbone. We also plot the loss curves of object classification L_{cls} and localization L_{bbox}, as shown in Figure 7. We find that MIM model helps localization task converge faster and better, and the supervised model benefits more for object classification. This also matches our previous observations, that MIM model can perform better on geometric and motion tasks, and on par or slightly worse on the tasks that its categories are sufficiently covered by ImageNet like COCO.

5. Related Work

Visual Pre-training. Throughout the deep learning era, supervised classification on ImageNet [14] has been the dominant pretraining task. It is found to deliver strong fine-tuning performance on numerous semantic understanding tasks [7, 17, 18, 23, 36, 38, 51, 53, 63, 65]. Over the past sev-

		Pose Estimation			Depth Estimation		Video Object Tracking		
backbone	pre-train	COCO *val*	COCO *test*	Crowd-Pose	NYUv2	KITTI	GOT10k *test*	Track-Net	LaSOT
SwinV2-B	1K-SUP	75.2	74.5	70.7	0.352	2.313	70.1	81.5	69.4
	22K-SUP	75.9	75.1	72.2	0.335	2.240	69.9	81.0	67.8
	1K-MIM	**77.6**	**76.7**	**74.9**	**0.304**	**2.050**	**70.8**	**82.0**	**70.0**
SwinV2-L	22K-SUP	76.5	75.7	72.7	0.334	2.150	71.1	81.5	69.2
	1K-MIM	**78.1**	**77.2**	**75.5**	**0.287**	**1.966**	**72.9**	**82.5**	**70.7**
Representative methods		HRFormer [76]			BinsFormer [46]		MixFormer [13]		
		77.2	76.2	72.5	0.330	2.098	75.6	83.9	70.1

Table 2. Comparisons of MIM and supervised (SUP) pre-trained models on the geometric and motion tasks. We report the AP (↑) for the pose estimation tasks, RMSE (↓) for the monocular depth estimation tasks, AO (↑) for the GOT10K dataset, and SUC (↑) for the TrackingNet dataset and LaSOT tracking dataset. The best results among the different pre-trained models are shown in the **bold** text. We provide the best results of the representative methods for reference.

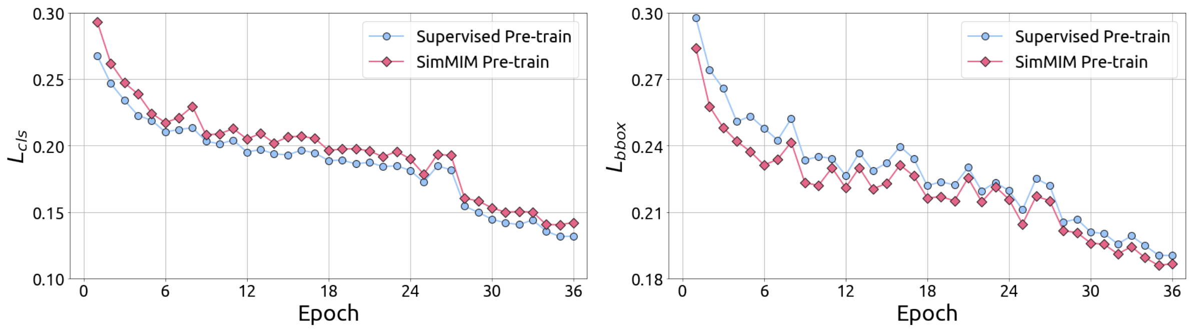

Figure 7. Loss curves of L_{cls} and L_{bbox} w.r.t the epoch number using supervised and MIM models with SwinV2-B as the backbone architecture.

eral years, self-supervised pretraining has attracted more and more attention, and achieved finetuning performance on par with the supervised counterparts on several representative downstream tasks [9, 28], including two representative ones, contrastive learning [5, 9, 19, 24, 28] and masked image modeling [3, 8, 27, 74]. In our work, we focus on understanding the different behaviors of supervised and emergent MIM pre-training.

Understanding Pre-training. There are some outstanding works [38,56–59,75,77] trying to understand the pre-training procedure and inspire a lot of following works in a wide range. [75] reveals how features of different layers are transferable in deep neural networks. [38] performs a sufficient experimental study on different backbones and tries to answer whether better ImageNet models transfer better. Some works [59,59,79] try to understand the behaviors of ViT, with CKA [37], loss landscape [43] and Fourier analysis. In NLP, after BERT [15] pre-training came out, there is also a lot of works [11, 25, 26, 39] trying to understand it. Most of them focus on the only interpretable component of Transformer, self-attention block, to give some detailed understanding.

6. Conclusion

In this work, we present a comprehensive analysis on masked image modeling, to reveal how and where MIM models work well. From visualizations, our most interesting finding is that the MIM pre-training brings locality to the trained model with sufficient diversity on the attention heads. This reveals why MIM is very helpful to the Vision Transformers (ViT, Swin, etc), because the Vision Transformer has a much larger receptive field, and to optimize it to a solution with strong generalization ability is difficult. In experiments, our most interesting finding is that MIM pre-training can perform very well on the geometric and motion tasks with weak semantics. This finding helps the model to achieve state-of-the-art performance on those benchmarks without bells and whistles.

Masked image modeling shows promise as a general-purpose pre-trained model. We hope that our paper encourages further exploration in the community and fosters new research in this direction. Ultimately, we aspire for this work to contribute to the understanding and motivation behind future technologies.

References

[1] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 456–473. Springer, 2022.

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

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[4] Likun Cai, Zhi Zhang, Yi Zhu, Li Zhang, Li Mu, and Xiangyang Xue. Bigdetection: A large-scale benchmark for improved object detector pre-training. *arXiv preprint arXiv:2203.13249*, 2022.

[5] Yue Cao, Zhenda Xie, Bin Liu, Yutong Lin, Zheng Zhang, and Han Hu. Parametric instance classification for unsupervised visual feature learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[8] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, and Heewoo Jun. Generative pretraining from pixels. *Advances in Neural Information Processing Systems*, 2020.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *ICML*, 2020.

[10] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[11] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[13] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. *arXiv preprint arXiv:2203.06717*, 2022.

[17] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655. PMLR, 2014.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[19] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in neural information processing systems*, pages 766–774, 2014.

[20] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019.

[21] Kunihiko Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological cybernetics*, 20(3):121–136, 1975.

[22] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[23] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[24] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[25] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Visualizing and understanding the effectiveness of bert. *arXiv preprint arXiv:1908.05620*, 2019.

[26] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. Self-attention attribution: Interpreting information interactions inside transformer. *arXiv preprint arXiv:2004.11207*, 2, 2020.

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020.

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[31] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016.

[32] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2021.

[33] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[34] Doyeon Kim, Woonghyun Ga, Pyunghwan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022.

[35] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[36] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning, 2019.

[37] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[38] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[39] Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*, 2019.

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

[41] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[42] Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pages 319–345. Springer, 1999.

[43] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

[44] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Haoshu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019.

[45] Yawei Li, Kai Zhang, Jiezhang Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *arXiv preprint arXiv:2104.05707*, 2021.

[46] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022.

[47] Liting Lin, Heng Fan, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *arXiv preprint arXiv:2112.00995*, 2021.

[48] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[50] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. *arXiv preprint arXiv:2111.09883*, 2021.

[51] 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.

[52] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer, 2021.

[53] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[54] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[55] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 310–327. Springer, 2018.

[56] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[57] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

[58] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[59] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34, 2021.

[60] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12159–12168, 2021.

[61] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[62] Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9629–9639, 2021.

[63] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[64] 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*, pages 746–760. Springer, 2012.

[65] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[66] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

[67] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[68] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[69] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[70] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021.

[71] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision*, pages 472–487. Springer, 2018.

[72] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection, 2021.

[73] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[74] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021.

[75] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[76] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. In *Advances in neural information processing systems*, pages 7281–7293, 2021.

[77] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

[78] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[79] Hong-Yu Zhou, Chixiang Lu, Sibei Yang, and Yizhou Yu. Convnets vs. transformers: Whose visual representations are more transferable? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2230–2238, 2021.

[80] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021.