Revealing the Dark Secrets of Masked Image Modeling

Zhenda Xie*13, Zigang Geng*23, Jingcheng Hu13, Zheng Zhang3, Han Hu3, Yue Cao3†
1Tsinghua University  2University of Science and Technology of China  3Microsoft 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 La-
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
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...
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
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 $224 \times 224$. 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 $4 \times 4$ resolution. The data augmentations include random flipping, half body transformation, random scale, random rotation, grid dropout, and color jittering. The input image size is $256 \times 256$ by default. We use the AdamW [54] optimizer with the base learning rate $5 \times 10^{-4}$ and the weight decay $5 \times 10^{-2}$. The learning rate is dropped to $5 \times 10^{-5}$ at the 120th epoch. We train the models for 150 epochs. We use a layer decay of $0.9^{0.9} / 0.85$ 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-
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.

<table>
<thead>
<tr>
<th>pre-train</th>
<th>Concept Generalization (CoG)</th>
<th>Kornlith et al’s 12 datasets (K12)</th>
<th>iNat18</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$L_1$</td>
<td>$L_2$</td>
<td>$L_3$</td>
</tr>
<tr>
<td>1K-SUP</td>
<td>79.4</td>
<td>76.2</td>
<td>72.7</td>
</tr>
<tr>
<td>1K-MIM</td>
<td>79.6</td>
<td>77.1</td>
<td>73.6</td>
</tr>
</tbody>
</table>

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.

following data augmentations: random horizontal flip, random brightness/gamma/hue/saturation/value and random vertical CutDepth. We randomly crop the images to $480 \times 480 / 352 \times 352$ 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 $224 \times 224$ and $112 \times 112$.

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-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 $384 \times 384$, MIM pre-trained SwinV2-L reaches 78.4 on COCO test-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-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 $320 \times 320$. We obtain the best SUC of 70.7 on the LaSOT with SwinV2-L backbone with the input image size $224 \times 224$ and template size $112 \times 112$.

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 ($1024 \times 1024$ 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-
several 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.

### 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 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.

<table>
<thead>
<tr>
<th>backbone</th>
<th>pre-train</th>
<th>Pose Estimation</th>
<th>Depth Estimation</th>
<th>Video Object Tracking</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>COCO val</td>
<td>COCO test</td>
<td>Crowd-Pose</td>
</tr>
<tr>
<td>SwinV2-B</td>
<td>1K-SUP</td>
<td>75.2</td>
<td>74.5</td>
<td>70.7</td>
</tr>
<tr>
<td></td>
<td>22K-SUP</td>
<td>75.9</td>
<td>75.1</td>
<td>72.2</td>
</tr>
<tr>
<td></td>
<td>1K-MIM</td>
<td><strong>77.6</strong></td>
<td><strong>76.7</strong></td>
<td><strong>74.9</strong></td>
</tr>
<tr>
<td>SwinV2-L</td>
<td>22K-SUP</td>
<td>76.5</td>
<td>75.7</td>
<td>72.7</td>
</tr>
<tr>
<td></td>
<td>1K-MIM</td>
<td><strong>78.1</strong></td>
<td><strong>77.2</strong></td>
<td><strong>75.5</strong></td>
</tr>
<tr>
<td>Representative methods</td>
<td>HRFormer [76]</td>
<td>77.2</td>
<td>76.2</td>
<td>72.5</td>
</tr>
</tbody>
</table>

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.

### 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


[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.


