Appendix

A. Broader related work

Self-supervised learning (SSL) of visual representations has lately been of great interest to the scientific community, opening up the possibility of learning powerful models without labeled data [1, 10]. SSL requires an appropriate *pretext task* which replaces a data-defined objective, and over the years, a plethora of such tasks have been proposed [28, 32, 44, 71], with Joint-Embedding Architectures (JEA) [15, 16, 18–21, 34, 36, 45, 69, 72], and Masked image modeling (MIM) [59, 60] gaining the most prominence in recent years.

Limitations of JEA models have been extensively covered by recent literature. JEA models rely on hand-crafted data augmentations [56], and their learned invariance to data perturbations can adversely affect the quality of representations [17, 41, 51, 63]. Moreover, JEA pretraining implicitly assumes a similar distribution of its pretraining and downstream task data [4], causing a need for additional dataset curation [45]. Therefore, development of SSL paradigms alternative to JEAs, including MIM, is an active line of research [5, 31, 58, 68].

Comparisons of Masked Image Modeling and Joint-Embedding Architectures have been the focus of several works, which tried to understand the differences and combine the advantages of both paradigms [9, 39, 47, 70]. The authors of [39, 70] frame MIM as a JEA that learns invariance to image occlusions, but find its representation to be less expressive than in other JEAs. A theoretical study of learning by reconstruction, conducted in [9], shows that data features required for reproducing pixels are misaligned with those needed for high-level perception. As a solution, multiple works propose shifting the prediction target from low-level pixels to higher-level image features, such as Histograms of Oriented Gradients [62] or latent representations [5, 50, 66], akin to the JEA objective. Finally, [47] thoroughly compare the properties of MIM and JEA-trained models including, similarly to us, the attention mechanisms of their patch tokens. They find that whereas JEAs form global and homogeneous attention maps, the attention of MIM patch tokens is more localized. Furthermore, [38, 65] show that MIM-pretrained transformers produce attention patterns that capture diverse image aspects, useful for tasks which require spatial understanding of images. Our work significantly extends these studies – we analyze the `[cls]` and patch representations of models trained with both paradigms and provide a detailed description of the information flow within them. We find that the attention mechanism emergent in MIM models imposes limitations that prevent these models from realizing their full potential in high-level perception tasks. Although this consequence of masked pretraining has previously been hinted at in the language models literature [30], to the best of our knowledge, it has not yet been discussed in the context of computer vision. While [30] address this with a modified pretraining scheme, we present Selective Aggregation as a lightweight solution for improving existing MIM representations without requiring architectural changes or additional pretraining.

B. Detailed experimental setup

In this section, we describe our experimental methodology: our choice of pretrained models, the details and hyperparameters of evaluating their representations, as well as the codebase used for the experiments.

B.1. Overview of the analyzed vision transformers

Our study aims to verify whether Selective Aggregation of patch token representations with AbMILP can yield form better representations than those of the `[cls]` tokens.

For this purpose, we analyze various vision transformer architectures that were pretrained with several MIM and JEA approaches, using the parameters shared by the authors of the respective methods. This has two advantages:

- Using the existing parameters significantly reduces the computational resources required for our study.
- Our study provides insights about the *very same* sets of parameters that are described in their respective literature and used by the wider research community.

For a fair evaluation, we use the parameters of the models that were pretrained on the ImageNet-1k dataset [54]. All of the explored model parameters are compatible with the implementations of the MAE [37] or SimMIM [64] vision transformers. Following the MAE and DINO implementations, when using ViT-S and ViT-B, we split the image into a 14×14 grid of patches of size 16×16. When using ViT-H, the we split the image into 16×16 patches of size 14×14.

The only analyzed models that are not publicly available but were trained by us are the ViT-S pretrained with the MAE and the fine-tuned ViT-S/B/L variants of the MAE. To prepare these models, we used the MAE pretraining and fine-tuning codebase and hyperparameters [37]. Before fine-tuning, we initialize the model with the pretrained MAE parameters as shared by the authors and use the `[cls]` token representation as input to the classifier.

B.2. Representation evaluation details

In our evaluation of ViT representations in terms of classification accuracy on ImageNet-1k and other large datasets (NUS-WIDE, COCO, Food-101), we follow the MAE linear probing protocol [37]: we augment the images only by random cropping, use the batch size of 16,384, and train the classifier head for 90 epochs (50 in the case of ViT-Large and Huge) with the LARS optimizer [33], the base learning rate of 0.1 with cosine decay and 10 epochs of warmup, optimizer momentum of 0.9, and no weight decay. For smaller datasets such as CUB, Stanford Cars, OxfordIIIPets, and ImageNet-1%, we follow a similar linear probing setup but train using SGD with a batch size of 1024. We report the results averaged over 3 random seeds. When using the AbMILP Selective Aggregation, we train it alongside the classifier head.

These evaluations are performed on a single node equipped with 4 NVIDIA-GH200 GPUs. Due to the memory constraints of this setup, we obtain the effective batch size of 16,384 by aggregating gradients from two forward passes with half of that batch size.

B.3. Codebase

Our code is based on the official MAE codebase [37], written in PyTorch [49], and available at `github.com/gmum/beyond_cls`. We include scripts required for the analysis of the attention mechanism in ViTs, as well as linear evaluation of their representations extended with AbMILP [43].

C. Additional experimental results

C.1. Analysis of information flow in self-supervised ViT architectures

This section contains the full details and experimental results of the attention mechanism in vision transformers, analyzed in Sec. 4. In the main manuscript, we include the analysis conducted on ViT-B, whereas in this section, we also provide the results of ViT-S and ViT-L architectures in Figures 10 to 13, For completeness, we re-include in them the pictograms describing each metric and the ViT-B results. We denote the contents of Figures 10 to 13 in Tab. 4. Due to the size of the figures, include them at the end of this supplementary material.

To complement the analysis, we report the average entropy of attention from the `[cls]` token to patch tokens across all Transformer blocks in Tab. 3. These results extend our findings beyond MAE and confirm that MIM models tend to distribute `[cls]` attention more uniformly across patch tokens compared to Joint-Embedding Architectures (JEAs). Notably, contemporary MIM architectures like I-JEPA and CAPI omit the `[cls]` token altogether, using average pooling over patch tokens instead. This results in an effective attention distribution that is equivalent to uniform and thus exhibits entropy values near the theoretical upper bound.

Detailed methodology. In our analysis, we aim to characterize the attention patterns resulting from MIM and JEA pretraining. Therefore, for both `[cls]` and patch tokens, we measure the attention of tokens to themselves (to see if tokens recycle their own information), and the entropy of attention to patch tokens (to see how information flows between the tokens).

Type	Masked Image Modeling					JEA
Model	MAE [37]	SimMIM [64]	BEIT-v2 [50]	I-JEPA [5]‡	CAPI [27]‡	DINO [16]
Aggr. entropy	5.03	4.96	4.89	5.28	5.28	4.70

‡ no `[cls]` token – uniform aggr.

Table 3. Patch aggregation entropy averaged across Transformer blocks in MIM models (we include DINO as a JEA representative for reference). MIMs aggregate patches more uniformly, motivating Selective Aggregation.

The entropy of an i-th token's attention to patch tokens (i.e. the $\mathbf{a}_{i,1:N}$ vector) is given by the Shannon entropy of its normalized values:

$$\mathbb{H}(\mathbf{a}'_i) = -\sum_{j=1}^{N} \mathbf{a}'_{i,j} \cdot log(\mathbf{a}'_{i,j}), \tag{6}$$

where $\mathbf{a}'_{i,1:N} = \frac{\mathbf{a}_{i,1:N}}{\sum_{j=1}^{N} \mathbf{a}_{i,j}}$. We measure these values for each self-attention head in each ViT block and report the average results per block. The inference is performed on the ImageNet-1k validation dataset (50,000 images).

To fairly compare Masked Image Modeling and Joint-Embedding paradigms, we analyze the ViT-B/16 models pretrained with MAE [37], DINO [16], MoCo-v3 [21], and iBOT [72], which represent prominent SSL approaches.[8] We use publicly available pretrained parameters provided by their respective authors. To examine whether optimizing for a global representation alters the attention behavior of MIM, we analyze an MAE model fine-tuned for ImageNet-1k classification using the `[cls]` token.

Analyzed models. As discussed in Appendix B.1, whenever possible, for each analyzed method, we use the ImageNet-1k pretrained model parameters officially released by their respective authors. The only exception to this is the MAE trained with ViT-S, which we trained ourselves, and the finetuned MAE (MAE-FT), which we finetuned ourselves for ImageNet-1k classification on top of the `[cls]` token features. Due to the lack of available ViT-L parameters of MoCo-v3 [21] and DINO [16], we omit them from the analysis of this architecture. However, given that the three JEA approaches behave similarly for each property analyzed in ViT-S and ViT-B architectures, we believe that the available ViT-L iBOT [72] variant sufficiently represents JEA. Similarly, we do not conduct this comparison with the ViT-H architecture, due to the lack of publicly available parameters of ViT-H trained with JEA to compare with.

[8] While iBOT optimizes a hybrid of JEA and MIM objectives, its performance gains are largely attributed to the JEA component [72], which is why we categorize it as such.

Metric	ViT-B results (manuscript)	ViT-S/B/L results (Appendix)
[cls]-[cls] attention	Fig. 3	Fig. 10
[cls]-patch entropy	Fig. 4	Fig. 11
patch-patch attention	Fig. 5	Fig. 12
patch-patch entropy	Fig. 6	Fig. 13

Table 4. A reference of Figures depicting the analysis of the attention mechanism and their extended counterparts in the Appendix.

Discussion. We are interested in the behavior of the ViT attention mechanism emergent in the MAE and JEA approaches, especially in the deep ViT blocks which form higher-level image representations [67]. Across the three ViT architectures analyzed, we observe several consistent trends, more generally discussed in Section 4 and summarized below:

- The [cls] token of pretrained and fine-tuned MAE assigns a large portion of attention (around 40-50%) to its own representation.
- The entropy of attention between the [cls] and patch tokens is much higher in MAE than in the rest of the models, indicating that it aggregates the information from a larger number of patch tokens. Fine-tuning of the MAE decreases this value to the levels observed in JEA models, increasing the selectiveness of attention.
- The attention of MAE patch tokens to themselves (relative to all patch tokens) is higher than in other models, indicating they are more likely to preserve their own, diverse information [47]. Fine-tuning of the MAE results in lowering this metric to the level observed in the JEA models. MAE patches also attend to to other patches with lower entropy than in JEAs and this does not change after fine-tuning.

C.2. Designing the token aggregation mechanism

In this section, we discuss different design choices for the token aggregation function, which uses either various variants of AbMILP [43], or other, non-trainable substitutes. Unless sepcified otherwise, all experiments reported in this section are conducted with the ViT-B model pretrained with the MAE [37].

Ablation study of AbMILP variants. We explore several designs of the model used by AbMILP to predict the scores for patch aggregation and report their performance in Tab. 5.

The original AbMILP architecture [43] uses a 2-layer MLP with the Tanh activation function. MAE patch tokens aggregated by this model achieve an accuracy of 68.7%. Although this is higher than the [cls] token representation, we found that the training process is unstable and replaced the Tanh activation with ReLU. This led to more stable training and an improvement in accuracy by almost 3 pp. Surprisingly, reducing the MLP to a single linear layer achieves almost the same results. Due to the simplicity and performance of this design, we adopt it in our main experiments. As seen in Sec. 5.1, the effectiveness of this approach generalizes to aggregating representations of MIM models other than the MAE.

Activation function	AbMILP MLP depth 1†	2	3	4
ReLU	71.6	71.7	71.5	71.6
GeLU	†linear model	71.6	71.6	71.5
Tanh	w/o activation	68.7	66.7	66.7

Table 5. Comparison of ImageNet-1k classification accuracy of the MAE representation aggregated by different variants of AbMILP [43]. Deeper MLPs do not boost performance.

We note that AbMILP is just one of several Multiple-Instance Learning methods that can be adopted to aggregate patch token representations. As an alternative, we explore the Self-Attention AbMILP [26] where, prior to computing the aggregation scores and the aggregated representation, tokens are processed by an additional trainable self-attention head. This approach achieves accuracy much closer to that of the JEA-trained approaches – 74.83%. This indicates an even larger richness of information stored in the representation space of Masked models, which requires more complex task-specific heads in order to be fully exploited. However, we found the training of this model to be unstable with the LARS optimizer [33], and were only able to train it using SGD. Moreover, a classification head that internally uses trainable self-attention to pre-process the classifier input is incomparable to a simple linear probe. For these reasons, we do not include this approach in our main experiments.

Non-trainable token aggregation. Apart from the AbMILP-based aggregation, we explore several alternative token aggregation functions that are not trained along with the classifier model. We discuss these approaches and their properties below and report their representations' average accuracies and entropies of the aggregation vectors in Tab. 6. To measure if different token aggregation approaches select the same patch tokens, in Fig. 7, we report the average Kullback-Leibler Divergence between token selection vectors produced by each method. Finally, we visualize the example token selection vectors in Fig. 9.

- **Average MAE [cls] token attention** – the average attention between the [cls] and patch tokens, produced by the MSA of the final MAE ViT block. As evidenced by the high entropy, this approach aggregates many patches, achieving quality similar to that of the regular [cls] representation.

- **Lowest-entropy MAE [cls] token attention** – the attention map between the [cls] and patch tokens produced by the MSA of the final MAE ViT block, which has the lowest entropy. This approach achieves low aggregation entropy, but due to the diversity of image fragments attended by different self-attention heads [47], the attended fragment of an image is not guaranteed to contain the object of interest.
- **MAE central patch token attention** – the average attention between the token of the central patch in the image and other patches. This approach can distinguish the tokens of the object of interest as long as it is depicted on the central image patch, which is not always the case. As evidenced by the high KLD between the Lowest-entropy MAE [cls] token attention and MAE central patch token attention, these two approaches tend to have a low agreement in terms of which tokens to select, suggesting their high volatility.
- **Average DINO [cls] token attention** – the average attention between the [cls] and patch tokens, produced by the MSA of the final DINO ViT block. As observed by [16], DINO attention maps are exceptionally good at capturing the main objects of interest in the images. MAE patch tokens selected with this approach form representations superior to the [cls] token, but an obvious drawback of this approach is the reliance on an externally pretrained model. As seen in Fig. 7, this selects tokens most similar to the AbMILP-based token aggregation.

Token aggregation approach	Accuracy	Entropy
Average MAE [cls] token attention	67.8	5.14
Lowest-entropy MAE [cls] token attention	66.3	4.77
MAE central patch token attention	65.2	4.70
Average DINO [cls] token attention	70.9	4.89
AbMILP	71.6	4.80

Table 6. Evaluation of different token aggregation approaches in terms of classification accuracy of their representations, and entropy of the aggregation vectors they produce.

Most of the above approaches select the MAE patch tokens with an entropy close to that observed in the JEA [cls] token. However, except for the attention maps generated by DINO and AbMILP, we did not find an approach that would reliably select patch tokens to form a representation of better quality than the [cls] token. Finding such tokens in an unsupervised manner is an interesting direction for future work.

Selective Aggregation and Attentive Probing Attentive Probing (AP) [22] has been proposed as an alternative to naive feature aggregation in ViTs. Similarly to our Selective Aggregation, AP learns to emphasize the most relevant patch tokens while keeping the encoder parameters frozen. However, AP differs from our approach in a key way: it does not only learn to aggregate tokens, but also transforms them with a cross-attention layer into a new representation space. potentially more suitable for the downstream task [9]. In contrast, AbMILP is designed to isolate the aggregation process while preserving the original ViT representations.

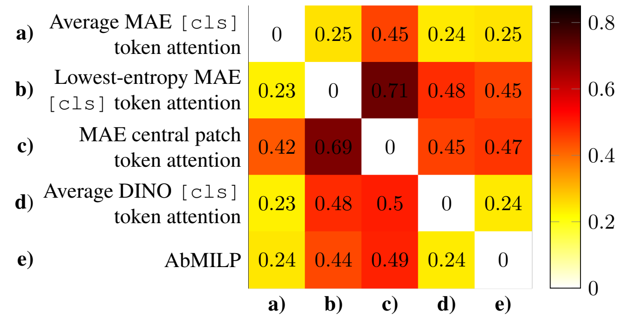

Figure 7. Mean KLD between aggregation vectors produced by different token aggregation techniques.

We compare AP and AbMILP across multiple MIM models in terms of ImageNet-1k classification and report the results in Tab. 7. Since AP typically uses a 12-head self-attention mechanism, we additionally evaluate a reduced variant with a single attention head (without reducing the representation dimensionality) to better compare with the capacity of AbMILP (which predicts a single set of representation aggregation weights). As expected, the full AP model achieves the best results, benefiting from its greater expressive power. However, despite AP's significantly higher parameter and compute cost, reducing it to a single head brings its performance in line with AbMILP. This result is somewhat surprising and suggests that AP's strength may come from ensembling multiple Selective Aggregation patterns rather than from the learned transformation. Exploring this insight to develop more efficient Selective Aggregation strategies is a promising direction for future work.

Encoder		Aggregation method		
Initialization	ViT type	AbMILP	AP (1 head)	AP (12 heads)
MAE [37]	ViT-S	54.4	53.6	**63.9**
MAE [37]	ViT-B	71.6	71.4	**75.4**
MAE [37]	ViT-L	77.4	77.6	**79.7**
MAE [37]	ViT-H	78.1	78.3	**80.0**
BEIT-v2 [50]	ViT-B	80.9	81.0	**81.8**
I-JEPA [5]	ViT-H	79.2	79.5	**79.7**
CAPI [27]	ViT-L	82.4	81.6	**82.7**

Table 7. Comparison of AbMILP [43] and Attentive Probing (AP) [22] aggregation schemes. AbMILP and the single-head cross-attention AP perform comparably.

Encoder		Localization based on	
Source	ViT	`[cls]` attention map	Selective Aggregation map
MAE [37]	ViT-B	53.3	**59.4**
BEIT-v2 [50]	ViT-B	44.3	**65.1**

Table 8. Object localization capabilities of the `[cls]` attention and Selective Aggregation weights, measured in terms of MaxBoxAccV2 [24] on the ImageNet validation dataset.

C.3. Using Selective Aggregation for object localization.

While global representations, which are the focus of this paper, are not generally suitable for dense prediction tasks, their attention maps can be used as a means to localize the object of interest in the image [16]. Because Selective Aggregation highlights the most relevant tokens, it can be used in a similar manner. We evaluate this capability of Selective Aggregation with the MAE and BEIT-v2 models, comparing it to their `[cls]` attention maps. We measure the localization quality in terms of MaxBoxAccV2 [24, 52] on the ImageNet validation dataset. We report the results in Tab. 8, and visualize the example results in Fig. 8. Our results indicate that the more focused Selective Aggregation localizes the objects of interest more accurately.

D. Future research directions

Our results indicate that lack of global representation aggregation is inherent to vision transformers trained with Masked Image Modeling. In this section, we summarize several potential research directions for better understanding this issue.

Unsupervised discovery of relevant tokens. We have showed that a shallow AbMILP [43] is sufficient for recognizing the patch tokens of MIM models that are relevant to form global image representations. However, in each MIM model, we learn that function together with the classifier dedicated to downstream tasks. Understanding what makes a patch token relevant for global representation and finding such tokens in an unsupervised manner is a natural further direction.

Scaling Selective Aggregation. Our implementation uses the minimal version of the aggregation score prediction model. In our comparison with Attentive Probing, we show that it succeeds not necessarily due to further processing of representations, but rather due to an ensemble of multiple self-attention heads. A full study of the effectiveness of vertical (more complex transformations) and horizontal (larger ensemble of aggregation functions) scaling of Selective Aggregation would be very beneficial for determining the most efficient MIM adaptation protocol.

Aggregation of internal ViT representations. Currently, Selective Aggregation acts only act on patch representations of the final ViT block. While this approach improves the MIM representations, we note that it does not interfere in any way with their internal information flow. However, as shown in Fig. 4, the `[cls]` token of JEAs aggregates patch information increasingly selectively throughout the several final model blocks. We hypothesize that similarly aggregating MIM representations within internal ViT blocks, either via additional training objectives or post-pretraining modifications, could yield further improvements in their quality.

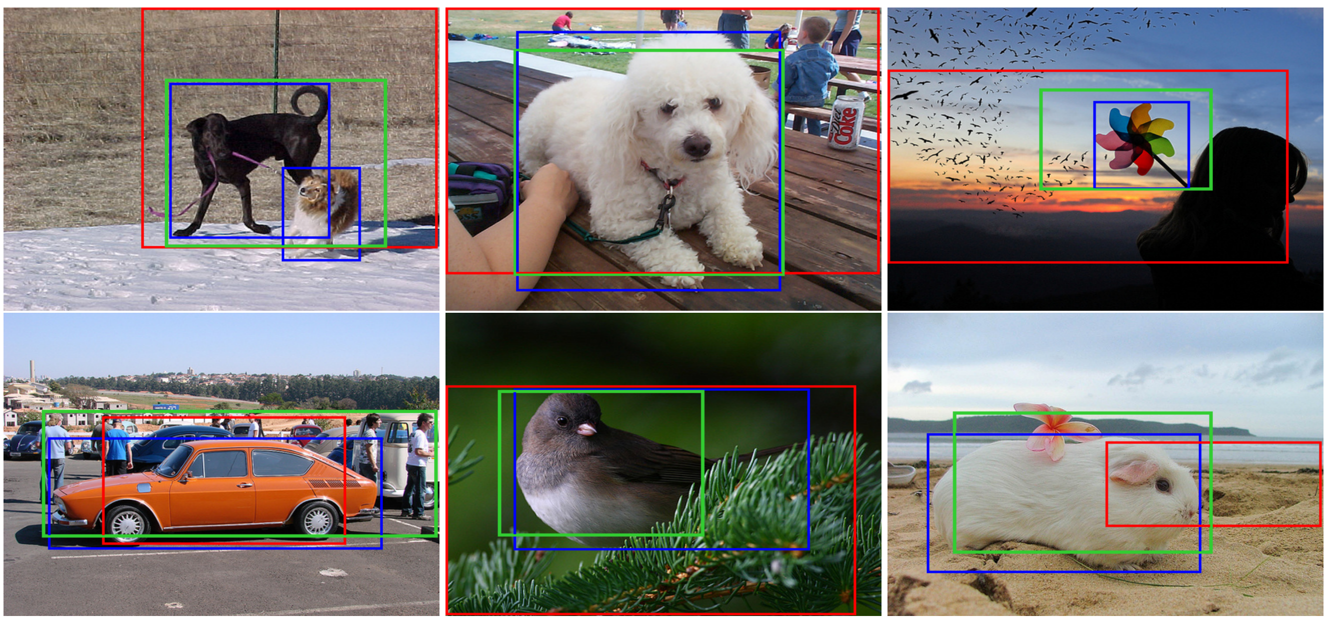

Figure 8. Example localization results of the MAE `[cls]` attention and Selective Aggregation weights. Blue: ground-truth. Red: bounding box predicted from the `[cls]` attention map. Green: bounding box predicted from the Selective Aggregation scores. Selective Aggregation locates objects with better accuracy (see Tab. 8).

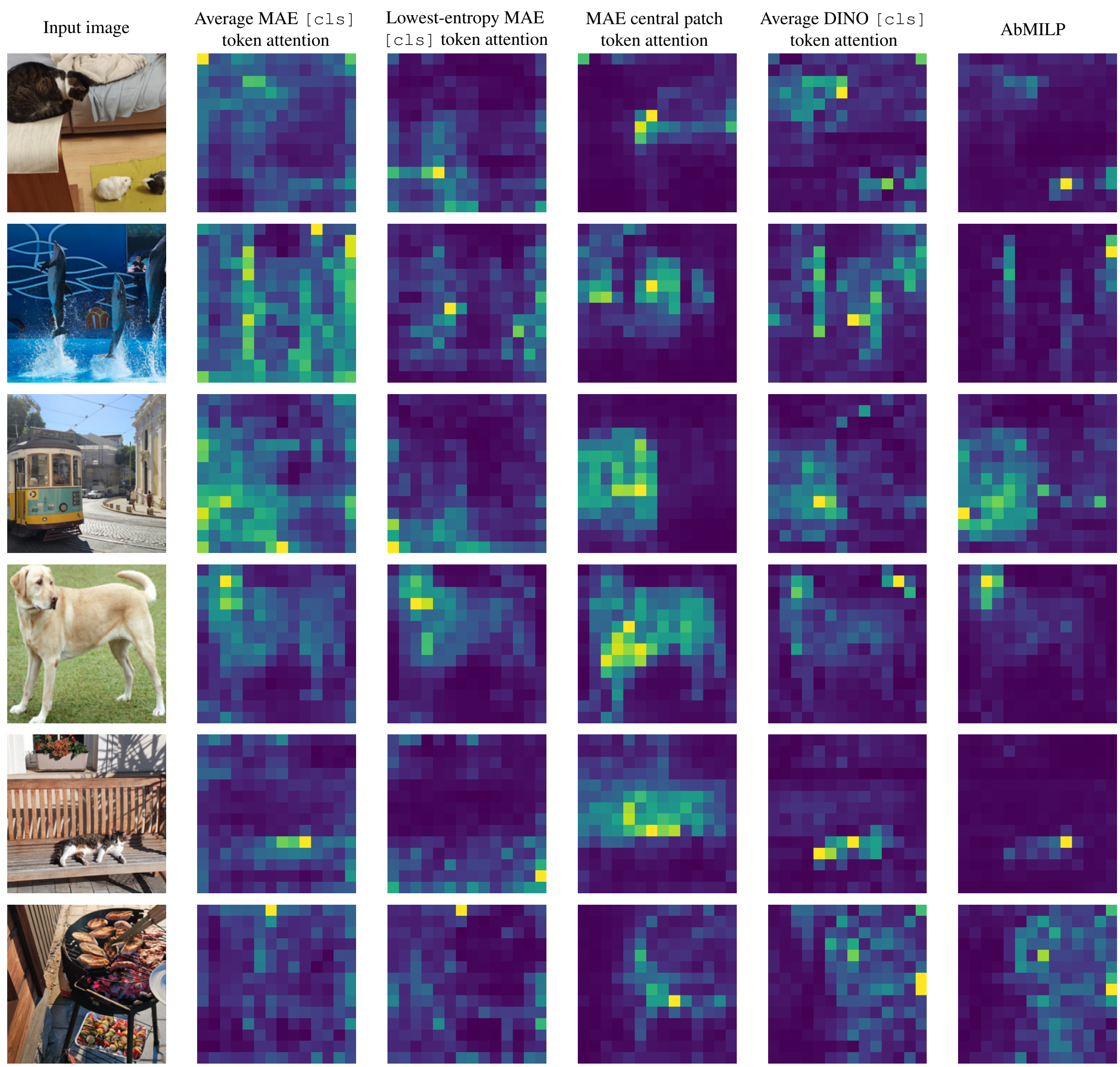

Figure 9. Example token aggregation scores produced by different approaches denoted in columns. The average [cls] attention of the MAE aggregates the patches too uniformly. The [cls] attention with lowest entropy and the attention of the central patch have low entropy, but are not guaranteed to capture the object of interest in the image. Finally, the DINO [cls] attention maps and aggregation vectors produced by AbMILP reliably identify the most crucial patches for forming high-level global image representations.

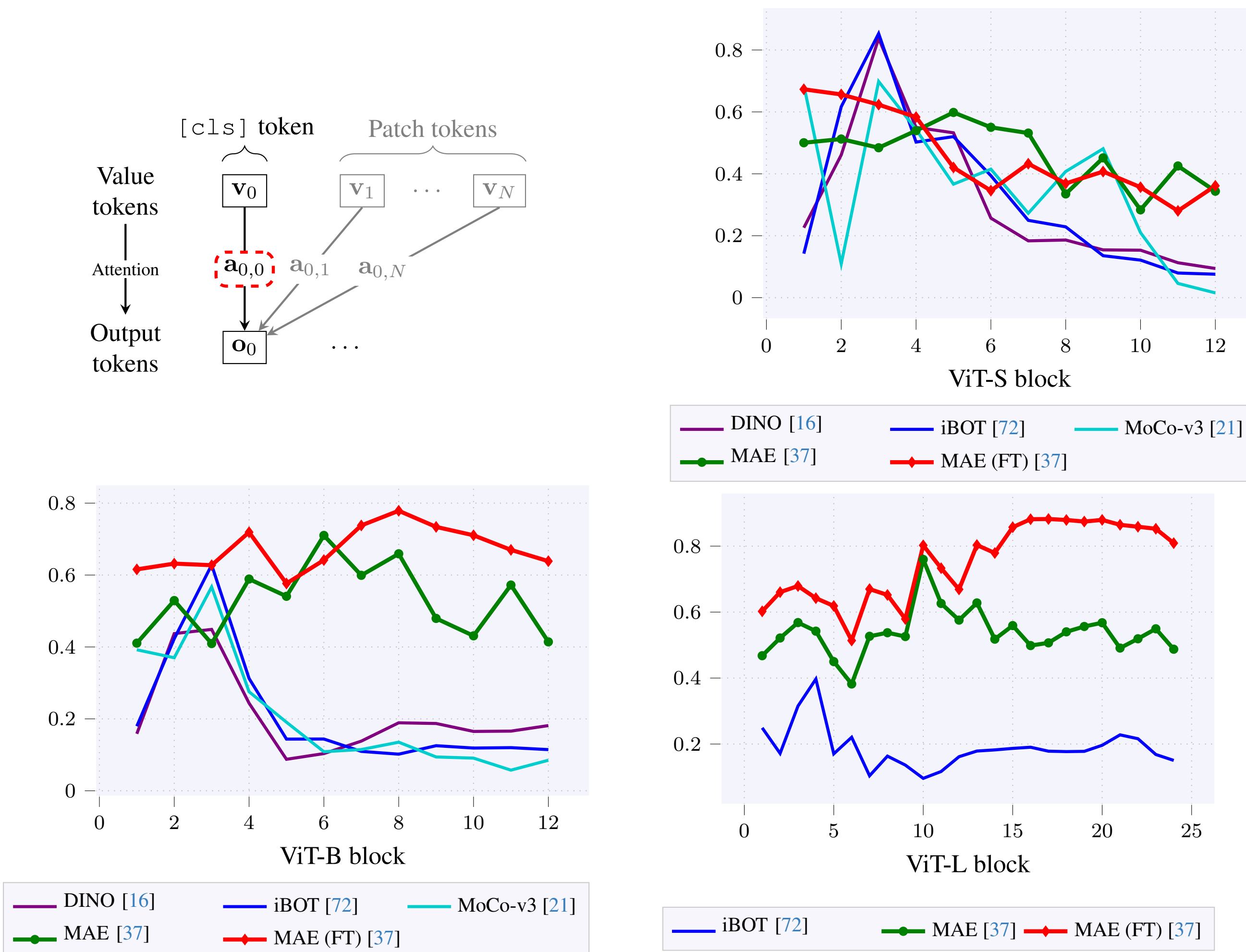

Figure 10. Extended version of Figure 3. Attention of the `[cls]` token to itself is much higher in both pretrained and finetuned MAE, than in the JEA ViTs. As opposed to JEA, where the `[cls]` tokens gather a large amount of information from the patch tokens, the MAE `[cls]` token primarily recycles its own representation.

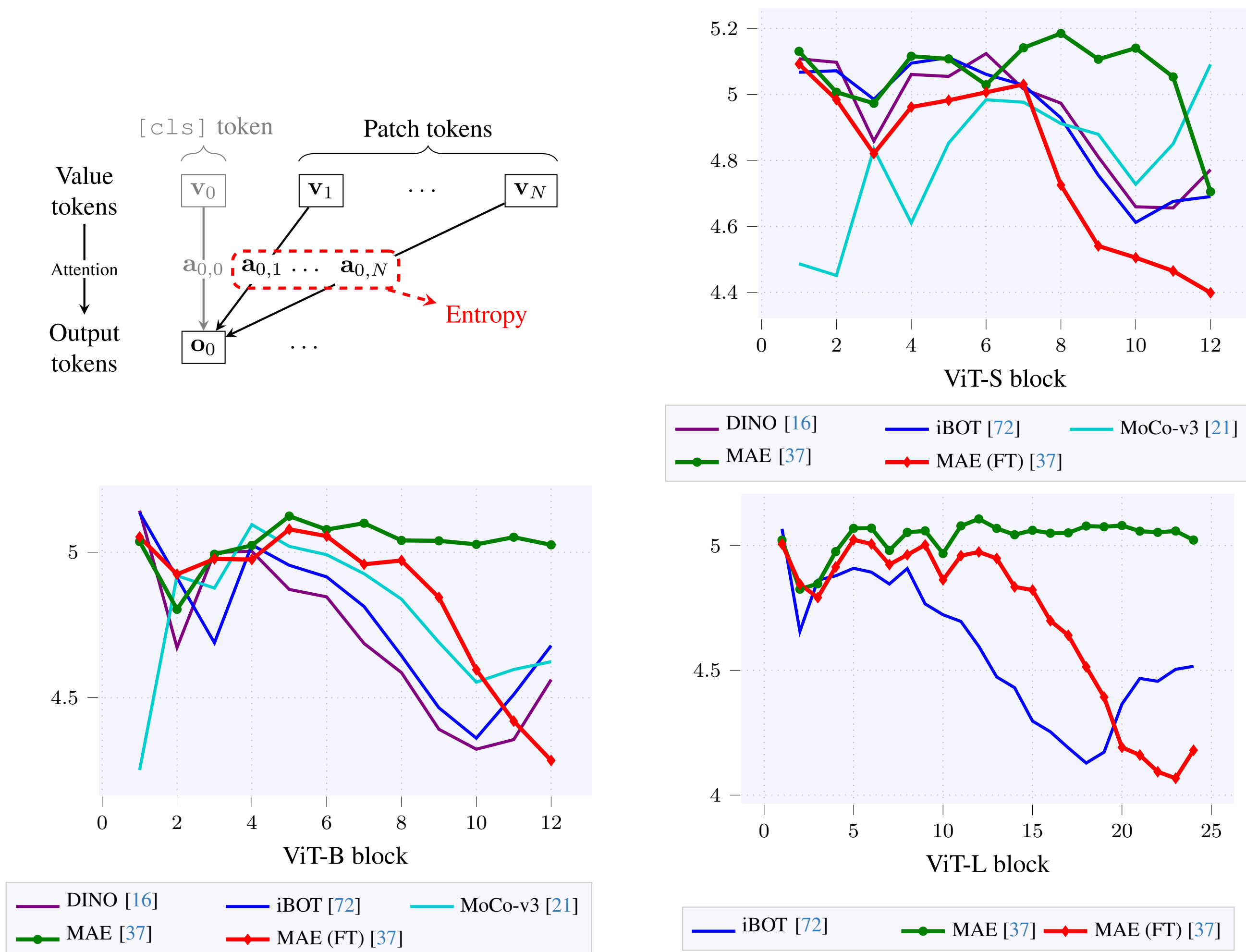

Figure 11. Extended version of Figure 4. Entropy of attention between the `[cls]` and patch tokens. In MAE, its value reaches almost the maximal possible level, In other models, it decreases in the deeper model blocks, indicating that the `[cls]` token attends to different patches in a more selective manner. Fine-tuning of MAE decreases this entropy. indicating that selective attention to patch tokens is crucial for good perception.

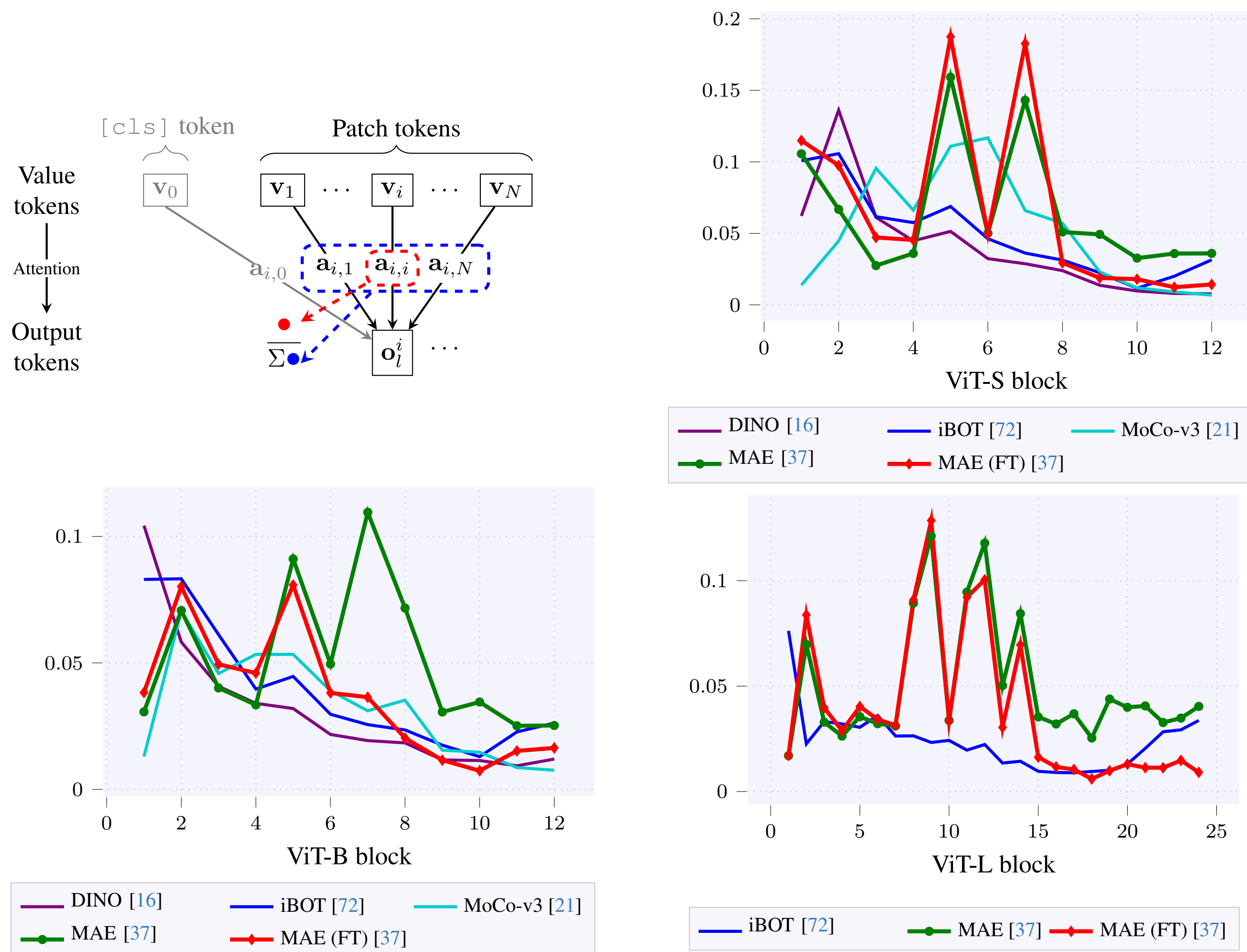

Figure 12. Extended version of Figure 5. Attention of the patch tokens to themselves, relative to the total attention assigned to all patch tokens. In the latter MAE blocks, patch tokens seem to assign the largest amount of relative attention to themselves, compared to the tokens of JEA.

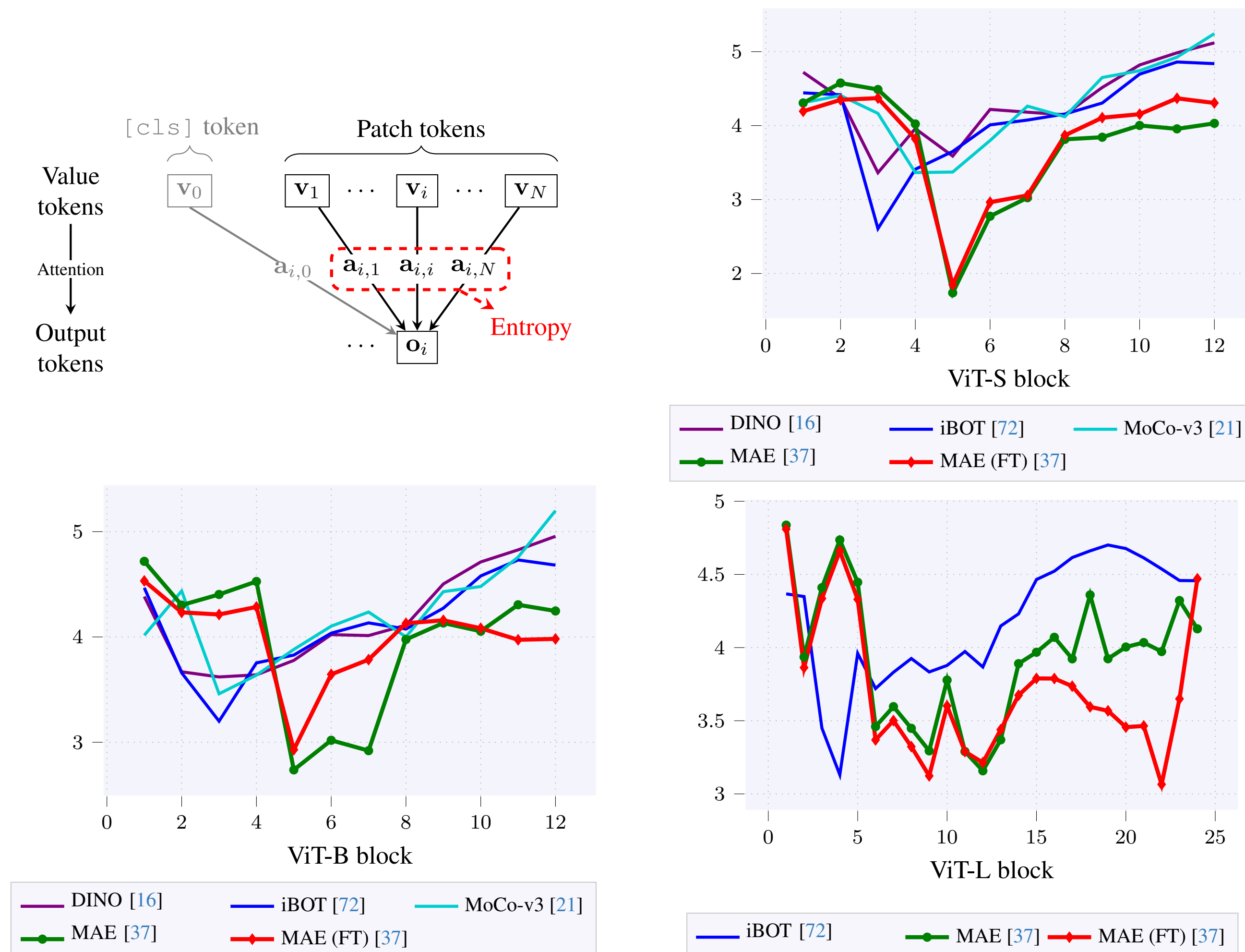

Figure 13. Extended version of Figure 6. Entropy of attention of patch tokens to patch tokens. In MAE, the patch tokens attend to other patches with lower entropy than in JEA, suggesting that they form a representation of their local image fragments.

References

[1] Saleh Albelwi. Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy*, 24(4), 2022. 2, 9

[2] Benedikt Alkin, Lukas Miklautz, Sepp Hochreiter, and Johannes Brandstetter. MIM-refiner: A contrastive learning boost from intermediate pre-trained masked image modeling representations. In *The Thirteenth International Conference on Learning Representations*, 2025. 7

[3] 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 *European Conference on Computer Vision*, pages 456–473. Springer, 2022. 3, 7

[4] Mido Assran, Randall Balestriero, Quentin Duval, Florian Bordes, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, and Nicolas Ballas. The hidden uniform cluster prior in self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 9

[5] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15619–15629, 2023. 1, 2, 4, 7, 8, 9, 10, 12

[6] Sara Atito, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *arXiv preprint arXiv:2104.03602*, 2021. 1, 2, 3

[7] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 3

[8] Bahdanau and others. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 6

[9] Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 4, 7, 9, 12

[10] Randall Balestriero, Mark Ibrahim, Vlad Sobal, Ari S. Morcos, Shashank Shekhar, Tom Goldstein, Florian Bordes, Adrien Bardes, Grégoire Mialon, Yuandong Tian, Avi Schwarzschild, Andrew Gordon Wilson, Jonas Geiping, Quentin Garrido, Pierre Fernandez, Amir Bar, Hamed Pirsiavash, Yann LeCun, and Micah Goldblum. A cookbook of self-supervised learning. *ArXiv*, abs/2304.12210, 2023. 1, 2, 9

[11] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers, 2022. 1, 4

[12] Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024. 3

[13] Florian Bordes, Randall Balestriero, Quentin Garrido, Adrien Bardes, and Pascal Vincent. Guillotine regularization: Why removing layers is needed to improve generalization in self-supervised learning. *Transactions on Machine Learning Research*, 2023. 3

[14] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 8

[15] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, pages 9912–9924. Curran Associates, Inc., 2020. 1, 2, 9

[16] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 16, 17, 18, 19

[17] Ruchika Chavhan, Jan Stuehmer, Calum Heggan, Mehrdad Yaghoobi, and Timothy Hospedales. Amortised invariance learning for contrastive self-supervision. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 9

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 9

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15750–15758, 2021.

[20] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020.

[21] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 1, 3, 4, 5, 6, 7, 9, 10, 16, 17, 18, 19

[22] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *Int. J. Comput. Vision*, 132(1):208–223, 2023. 3, 12

[23] Xuweiyi Chen, Markus Marks, and Zezhou Cheng. Probing the mid-level vision capabilities of self-supervised learning, 2024. 1

[24] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. to appear. 13

[25] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, New York, NY, USA, 2009. Association for Computing Machinery. 8

[26] Rymarczyk D., Borowa A., Tabor J., and Zielinski B. Kernel self-attention for weakly-supervised image classification using deep multiple instance learning. In *WACV*, 2021. 11

[27] Timothée Darcet, Federico Baldassarre, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Cluster and predict latents patches for improved masked image modeling, 2025. 1, 2, 4, 7, 8, 10, 12

[28] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 9

[29] 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. 2, 3, 7

[30] Luyu Gao and Jamie Callan. Condenser: a pre-training architecture for dense retrieval, 2021. 9

[31] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning, 2024. 2, 9

[32] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 9

[33] Boris Ginsburg, Igor Gitman, and Yang You. Large batch training of convolutional networks with layer-wise adaptive rate scaling, 2018. 10, 11

[34] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284. Curran Associates, Inc., 2020. 1, 9

[35] 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 (CVPR)*, 2016. 3

[36] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 9

[37] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16000–16009, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 16, 17, 18, 19

[38] Yu Huang, Zixin Wen, Yuejie Chi, and Yingbin Liang. How transformers learn diverse attention correlations in masked vision pretraining. In *ICML 2024 Workshop on Theoretical Foundations of Foundation Models*, 2024. 2, 9

[39] Xiangwen Kong and Xiangyu Zhang. Understanding masked image modeling via learning occlusion invariant feature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6241–6251, 2023. 2, 9

[40] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 8

[41] Hankook Lee, Kibok Lee, Kimin Lee, Honglak Lee, and Jinwoo Shin. Improving transferability of representations via augmentation-aware self-supervision. In *Advances in Neural Information Processing Systems*, pages 17710–17722. Curran Associates, Inc., 2021. 2, 9

[42] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 8

[43] Ilse M. et al. Attention-based deep multiple instance learning. In *ICML*, 2018. 2, 6, 10, 11, 12, 13

[44] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision – ECCV 2016*, pages 69–84, Cham, 2016. Springer International Publishing. 9

[45] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 1, 2, 7, 9

[46] Adam Pardyl, Grzegorz Rypeść, Grzegorz Kurzejamski, Bartosz Zieliński, and Tomasz Trzciński. Active visual exploration based on attention-map entropy. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 1303–1311. International Joint Conferences on Artificial Intelligence Organization, 2023. Main Track. 1, 2

[47] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 5, 6, 9, 11, 12

[48] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 8

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 10

[50] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers, 2022. 1, 2, 4, 7, 8, 9, 10, 12, 13

[51] Marcin Przewięźlikowski, Mateusz Pyla, Bartosz Zieliński, Bartłomiej Twardowski, Jacek Tabor, and Marek Śmieja. Augmentation-aware self-supervised learning with conditioned projector. *Knowledge-Based Systems*, 305:112572, 2024. 1, 2, 9

[52] Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, and Yannis Avrithis. Keep it simpool: Who said supervised transformers suffer from attention deficit? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5350–5360, 2023. 13

[53] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 2014. 7, 8, 9

[55] Jeongwoo Shin, Inseo Lee, Junho Lee, and Joonseok Lee. Self-guided masked autoencoder. In *Advances in Neural Information Processing Systems*, pages 58929–58954. Curran Associates, Inc., 2024. 6

[56] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? 33:6827–6839, 2020. 1, 2, 9

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 3

[58] Shashanka Venkataramanan, Mamshad Nayeem Rizve, Joao Carreira, Yuki M Asano, and Yannis Avrithis. Is imagenet worth 1 video? learning strong image encoders from 1 long unlabelled video. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 9

[59] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. 2, 3, 9

[60] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 2, 3, 9

[61] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 7, 8

[62] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training, 2023. 4, 7, 9

[63] Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021. 1, 2, 9

[64] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9653–9663, 2022. 1, 2, 3, 4, 7, 8, 9, 10

[65] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14475–14485, 2023. 2, 9

[66] xingbin liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4, 9

[67] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 11

[68] Yike Yuan, Huanzhang Dou, Fengjun Guo, and Xi Li. SemanticMIM: Marring masked image modeling with semantics compression for general visual representation, 2025. 9

[69] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1, 2, 9

[70] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. In *Advances in Neural Information Processing Systems*, pages 27127–27139. Curran Associates, Inc., 2022. 1, 2, 3, 9

[71] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *Computer Vision – ECCV 2016*, pages 649–666, Cham, 2016. Springer International Publishing. 9

[72] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image BERT pre-training with online tokenizer. In *International Conference on Learning Representations*, 2022. 1, 4, 5, 6, 7, 9, 10, 16, 17, 18, 19

[73] Jiaxin Zhuang, Linshan Wu, Qiong Wang, Varut Vardhanabhuti, Lin Luo, and Hao Chen. Mim: Mask in mask self-supervised pre-training for 3d medical image analysis, 2024. 1, 7