Stare at What You See: Masked Image Modeling without Reconstruction

Hongwei Xue[1,2]*, Peng Gao[2,3]†, Hongyang Li[2]†, Yu Qiao[2], Hao Sun[4], Houqiang Li[1], Jiebo Luo[5]

[1]University of Science and Technology of China [2]Shanghai Artificial Intelligence Laboratory
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[4] China Telecom Corporation Ltd. Data&AI Technology Company [5]University of Rochester

Abstract

Masked Autoencoders (MAE) have been prevailing paradigms for large-scale vision representation pre-training. By reconstructing masked image patches from a small portion of visible image regions, MAE forces the model to infer semantic correlation within an image. Recently, some approaches apply semantic-rich teacher models to extract image features as the reconstruction target, leading to better performance. However, unlike the low-level features such as pixel values, we argue the features extracted by powerful teacher models already encode rich semantic correlation across regions in an intact image. This raises one question: is reconstruction necessary in Masked Image Modeling (MIM) with a teacher model? In this paper, we propose an efficient MIM paradigm named MaskAlign. MaskAlign simply learns the consistency of visible patch features extracted by the student model and intact image features extracted by the teacher model. To further advance the performance and tackle the problem of input inconsistency between the student and teacher model, we propose a Dynamic Alignment (DA) module to apply learnable alignment. Our experimental results demonstrate that masked modeling does not lose effectiveness even without reconstruction on masked regions. Combined with Dynamic Alignment, MaskAlign can achieve state-of-the-art performance with much higher efficiency. Code and models will be available at https://github.com/OpenPerceptionX/maskalign.

1. Introduction

In recent years, Vision Transformers are showing tremendous potential in computer vision area [8, 40, 41]. Following the big success of masked modeling in the natural language processing [24], Masked Image Modeling (MIM) has demonstrated a great ability of self-supervised learning [3, 13], while alleviating the data-hungry issue of Transformer architectures. The visual representation learned through MIM shows promising performance on various downstream vision tasks, outperforming the contrastive learning paradigms [4, 6].

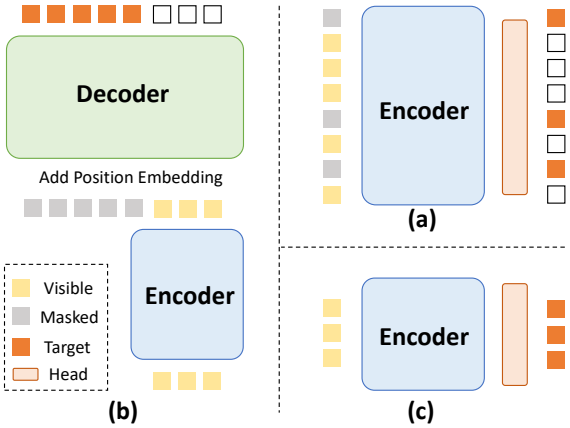

Figure 1. Comparison with existing paradigms of masked image modeling. **(a) Inpainting-style**: BEiT [3], MaskFeat [43], MVP [44], BEiT V2 [33], *etc*. They take the whole image with some mask token replacement as the input of Encoder. Then a Linear head is applied to predict masked feature. **(b) Decoder-style**: MAE [13], CAE [5], MCMAE [11], *etc*. They drop most tokens and take the rest as the input of Encoder. Then a multi-layer Transformer is applied to decode masked features from visible tokens. **(c) Ours.**: Our paradigm take some visible tokens as the input of Encoder and align visible tokens with target features *only*.

Existing Masked Image Modeling (MIM) methods aim to hallucinate the intact image from a small portion of visible image regions. As depicted in Fig. 1, existing MIM methods are mainly divided into two types: (a) inpainting-style [3, 43, 46] and (b) decoder-style [5, 11, 13]. These two types both require the model to reconstruct masked regions. The inpainting-style models replace image regions with learnable vectors then fill them by the interaction within the encoder. The decoder-style models drop image regions then decode features from masked regions' positions based on

*This work was performed when Hongwei Xue was visiting Shanghai AI Laboratory as a research intern.
†Corresponding authors.

the visible information. Some very recent works introduce semantic-rich teacher models like CLIP [35] into the two paradigms by using features extracted by teacher models as the reconstruction target [17, 33, 34, 44]. In light of the semantic knowledge learned by teacher models, these works further improve the representation after masked image modeling, leading to better performance.

Reconstruction on masked regions implicitly forces the model's encoder to understand the semantic correlations within an image. However, the reconstruction manner brings much computation on masked tokens within or outside the encoder in inpainting-style or decoder-style, respectively. This redundant computation decreases the training efficiency of the encoder thus increasing the pre-training cost. Unlike low-level and isolated features such as normalized pixel values of patches, Histogram of Oriented Gradients (HOG), etc., the feature map extracted by powerful teacher models already contains rich semantic correlations, learned during the teacher model training stage. This difference raises one question: is reconstruction the only way in Masked Image Modeling (MIM) with teacher models? To answer this question, we propose a much more efficient MIM paradigm named **MaskAlign** without any reconstruction on masked tokens.

On contrary of applying reconstruction on masked tokens, MaskAlign simply aligns the visible features extracted by the student model and intact image features extracted by the teacher model. As a consequence, MaskAlign forces the student model to learn not only good representation of the teacher model by feature alignment, but also the ability to hallucinate by masked modeling: feature consistency between the intact image and mask view requires the student model to infer semantics from much less information than teacher model. We adopt multi-level features of the teacher model as supervision to borrow richer semantics. However, the input of the student model contains much less information than the teacher model's, leading to misalignment of each layer's features. To tackle this problem, we enhance the student's features with a Dynamic Alignment (DA) module. DA dynamically aggregates different levels of student features and aligns with multi-level features of the teacher model. This approach can also easily transfer to asymmetric student-teacher structures.

From our experimental results, MaskAlign with a wide range of mask ratio outperforms the mask ratio of 0%, where it degenerates into Feature Distillation [45]. This verifies that masked modeling is still necessary for our paradigm. Meanwhile, our experiments validate the effectiveness of Dynamic Alignment by comparing different alignment strategies and numbers of feature levels. The combination of masked modeling and Dynamic Alignment makes our model achieve state-of-the-art results with much higher efficiency. For example, our model outperforms BEiT v2 [33] by 0.4% on ImageNet Finetuning Accuracy (from 85.0% to 85.4%) with 1/3 pre-training time only.

To sum up, our work has three-fold contributions:

1. We categorize and rethink existing Masked Image Modeling (MIM) paradigms and propose a more efficient MIM approach called MaskAlign. Even *without* any reconstruction on masked tokens, MaskAlign achieves new state-of-the-art performance with much higher efficiency.

2. We propose a Dynamic Alignment (DA) module to tackle the problem of input inconsistency between the student and teacher model, with negligible additional parameters and computation.

3. We conduct extensive experiments to verify the effectiveness of MaskAlign and Dynamic Alignment. Besides, our model shows a good ability of generalization on downstream tasks and larger size models.

2. Related Work

Masked Image Modeling. Motivated by Masked Language Modeling (MLM) in BERT [24], BEiT [3] explores Masked Image Modeling (MIM) on vision transformers by reconstructing the dVAE [37] feature extracted by DALL-E [36]. MAE [13] and SimMIM [46] find that RGB values can act as a simple yet good enough reconstruction target for masked modeling. PeCo [7], iBOT [52] and MaskFeat [6] respectively use dVAE with perceptual loss, an online tokenizer and the manually-crafted HOG descriptor, proving that the reconstruction target shows big impact. Motivated by this, MVP [44] firstly introduces multimodality-guided teacher models into MIM, by simply replacing the reconstruction target with CLIP [35] features. Rich-semantic guidance leads to impressive gains. Some very recent works: BEiT V2 [33], MILAN [17] and MaskDistill [34] also include CLIP in their model. BEiT V2 adopts CLIP features to train their discrete tokenizer, while MILAN and MaskDistill directly use CLIP features as reconstruction targets. All existing MIM works are based on reconstruction. In this paper, we explore a new paradigm of MIM without reconstruction, which significantly alleviates the efficiency issue brought by redundant computation of reconstruction.

Vision Language Pre-training. Learning visual linguistic representations from large-scale data has demonstrated unprecedented power in cross-modal learning [10, 12, 18, 21, 35, 39, 47, 49, 51]. Under the guidance of diverse texts with rich semantics, CLIP [35] advances the transfer performance on many downstream vision tasks, especially image-text generation [31, 32], requiring a sufficient understanding of semantic correlations within an image. Some existing

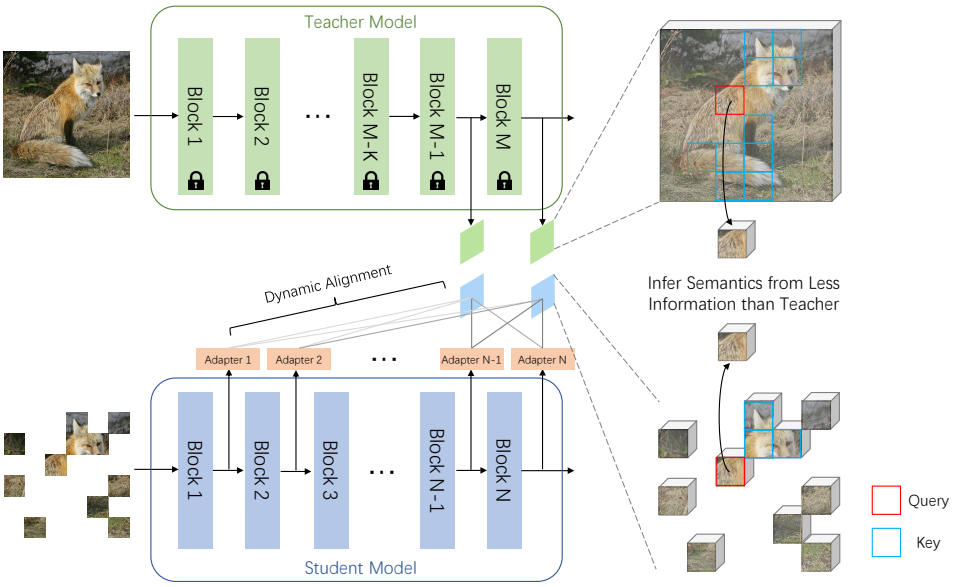

Figure 2. Framework of **MaskAlign**. MaskAlign aligns the visible features extracted by the student model and intact image features extracted by a frozen teacher model. Dynamic Alignment (DA) module learnably aggregates different level of student features and aligns with multi-level features of the teacher model.

masked modeling methods adopt the CLIP feature as the reconstruction target [17, 33, 44], outperforming counterparts using low-level features [6, 13]. In this work, we also adopt a frozen CLIP model to leverage good semantics and further improve the representation ability by incorporating masked image modeling.

Knowledge Distillation. Knowledge distillation (KD) [16] generates a soft label by the output of the teacher model to train the student model. KD transfers the capacity of teacher models into students and brings impressive gains. From that, KD has shown great potential in various tasks [15, 40, 50] and domains [22, 42]. Feature Distillation (FD) [45] finds that using the normalized dense feature from the teacher model to supervise student models can significantly advance the performances. In this paper, we leverage mask modeling in aligning with the teacher model instead of full-size input, leading to significant improvement in both performance and training efficiency.

3. Approach

MaskAlign aligns the visible features extracted by the student model and intact image features extracted by a frozen teacher model. The overview of MaskAlign is depicted in Fig. 2. In this section, we elaborate on the details of masking and alignment.

3.1. Model Structure

MaskAlign consists of a randomly initialized student model and a pre-trained teacher model. For the student model, we adopt standard Vision Transformer (ViT) as in [8] to make a fair comparison with existing works. We apply a frozen teacher model with rich semantics to produce the supervision. In experiments, we adopt ViT teacher models such as CLIP-ViT [35] and DINO [4]. For the input of teacher models, an image $I \in \mathbb{R}^{C\times H\times W}$ is divided to $N = HW/P^2$ patches: $\mathcal{I} = \{\boldsymbol{x}_i^p\}_{i=1}^N$ and $\mathcal{I} \in \mathbb{R}^{N\times\left(P^2C\right)}$, where (P, P) is the patch size. The image patches $\mathcal{I}$ are then linearly projected to input tokens with added positional embeddings. For the input of student models, the process is similar except we use a masked view of the image. Like in MAE, we drop $r\%$ patches and only feed visible patches $\mathcal{V} = \{\boldsymbol{x}_i^p\}_{i=1}^{N(1-r\%)}$ into student model. By Self-Attention mechanism within the Transformer, patches interact with others to aggregate information. In the student model, each patch (query) can only attend to $N(1-r\%)$ patches (keys), which are much less than the teacher model. This can greatly decrease the training cost and encourage the student model to learn a better ability of visual representation.

3.2. Masking Strategy

To eliminate the redundancy of an image, masking creates a task that cannot be easily solved by extrapolation from visible neighboring patches. To generate a mask view $\mathcal{V}$ from an intact image $\mathcal{I}$, one straightforward sampling strategy is random masking, which samples patches without replacement, following a uniform distribution [13].

Another masking strategy is based on the guidance of the teacher model. Following [17, 23, 48, 51], we also study attentive masking in our paradigm. Attentive masking aims

to feed tokens covering important image regions into the encoder with high probabilities. By doing so, the latent representations from the encoder provide sufficient information to infer semantics. We make a comparison of these two masking strategies in Sec. 4.

3.3. Dynamic Alignment

To borrow richer semantics from the teacher model, we use multi-level features as supervision. However, aligning student model's features with multi-level supervision has a challenge: as the input of the student model contains much less information than the teacher's, the input inconsistency causes misalignment between the student and teacher model on each layer. To tackle this problem, we propose a Dynamic Alignment (DA) module. DA can dynamically learn how to make alignment between the student and teacher model. A Transformer usually consists of a sequence of blocks. We add one adaptor A_i to each block's output x_i to project the student model's feature space to the teacher's. The adaptor could be a light model like a Linear layer or 2-layer MLP. To dynamically aggregate different levels of student features and align with multi-level features of the teacher model, we apply a Dynamic Alignment Matrix: W, which is an $S \times T$ matrix with entries w_{ij}, where S and T is the number of blocks in student and teacher model. The whole DA module can be formulated as:

$$\hat{y} = \{\sum_{i=0}^{S} w_{ij} A_i(x_i)\}_{j=0}^{T}. \tag{1}$$

where $\hat{y}$ is a set of linear combinations of multi-level features of the student model. During pre-training, gradients can backpropagate to Dynamic Alignment Matrix. In downstream tasks, we simply abandon the whole DA module.

To restrain the feature magnitudes of teacher features, we generate the alignment target $\tilde{y}$ by normalizing each level of teacher features as MAE [13] does on pixel values:

$$\tilde{y} = \{\text{Normalize}(y_i)\}_{j=0}^{T}. \tag{2}$$

Finally, following [45], we employ a smooth L1 loss between the student and teacher features:

$$\mathcal{L}_{\text{Align}}(\hat{y}, \tilde{y}) = \begin{cases} \frac{1}{2}(\hat{y} - \tilde{y})^2, & |\hat{y} - \tilde{y}| \leq 1 \\ (|\hat{y} - \tilde{y}| - \frac{1}{2}), & \text{otherwise} \end{cases}. \tag{3}$$

In experiments, we align student features with the latter layers of the teacher model, determined by a hyperparameter K. We also compare with alignment without a Dynamic Alignment Matrix, namely Layer-wise alignment. Layer-wise denotes aligning features layer-by-layer without a Dynamic Alignment Matrix. Experimental results in Sec. 4 demonstrate that Dynamic Alignment outperforms simple Layer-wise alignment, with nearly no increase in computational effort.

3.4. Relation to Existing Models

As depicted in Fig. 1, in this work we explore a new paradigm for masked image modeling. Inpainting-style models like BEiT V1/V2 [3, 33], MaskFeat [6], MVP [44] simultaneously process the partially masked image content and produce predictions for the masked patches. The full-size input leads to large computational costs for this kind of model. Decoder-style models like MAE [13], CAE [5] and MCMAE [11] only take the partial image as the input. The encoder maps the input into a latent representation, and the decoder reconstructs the input from the latent representation. The decoder learns the interaction on full-size features but is abandoned after pre-training. Both of these paradigms have the redundant computation of reconstruction. On the contrary of them, our model MaskAlign does not include any reconstruction on masked tokens. Our model only applies alignment on visible features extracted by the student model and intact image features extracted by the teacher model. As a result, our model has big advantages in efficiency and simplicity.

4. Experiments

4.1. Implementation

Pre-training. We pre-train MaskAlign on ImageNet-1k dataset [38], containing 1.28M training images. For the student model, we mainly study ViT-B/16 (12 blocks, 768 hidden size) and ViT-L/16 (24 blocks, 1024 hidden size). We employ the default 16×16 input patch size, partitioning the image of 224×224 into 14×14 patches. We only use standard random cropping and horizontal flipping for data augmentation, following MAE [13], CAE [5], etc. For pre-training, we train all models using AdamW optimizer [25], with a base learning rate of 1.5e-4, weight decay of 0.05, and optimizer momentum β_1, β_2 = 0.9, 0.95. We use a total batch size of 1024, and pre-train models under a cosine learning rate decay schedule with 10% warm-up steps. We also employ stochastic depth [19] with a 0.1 rate, and disable dropout in the Linear layer and Self-Attention.

Image Classification. After pre-training, we evaluate the model by fine-tuning on ImageNet-1K and report the top-1 accuracy on the validation set. We fine-tune our model 100 epochs with 5 warm-up epochs. We use the same batch size, optimizer, and weight decay as in pre-training. The base learning rate, layer-wise learning rate decay, and drop path rate are set to be 3e-4, 0.6 and 0.2, respectively. We use the Data Augmentation in MAE without any modification.

COCO Detection and Instance Segmentation. We also evaluate on COCO dataset [29] for object detection and instance segmentation to verify the transferability of our

Method	Backbone	Supervision	Forward Ratio	Reconstruct Ratio	Epochs	FT Acc.(%)
SimMIM [46]	Swin-B	RGB	100%	60%	800	84.0
MCMAE [11]	CViT-B	RGB	25%	75%	1600	85.0
MixMIM [30]	MixMIM-B	RGB	100%	100%	600	85.1
CMAE [20]	CViT-B	RGB	25%	75%	1600	85.3
BEiT [3]	ViT-B	DALLE	100%	40%	800	83.2
MAE [13]	ViT-B	RGB	25%	75%	1600	83.6
CAE [5]	ViT-B	DALLE	25%	75%	800	83.6
MaskFeat [43]	ViT-B	HOG	100%	40%	300	83.6
DMAE [2]	ViT-B	MAE-L	25%	75%	100	84.0
data2vec [1]	ViT-B	EMA	100%	60%	800	84.2
MVP [44]	ViT-B	CLIP-B	100%	40%	300	84.4
BEiT V2 [33]	ViT-B	CLIP-B	100%	40%	300	85.0
MaskDistill [34]	ViT-B	CLIP-B	100%	40%	300	85.0
MILAN [17]	ViT-B	CLIP-B	25%	100%	400	**85.4**
FD-CLIP [45]	ViT-B	CLIP-B	100%	**0%**	300	84.9
Ours	ViT-B	CLIP-B	30%	**0%**	200	**85.4**

Table 1. **Image classification** by fine-tuning on ImageNet-1K [38]. Our model achieves state-of-the-art performance with much fewer epochs. 'Forward Ratio' denotes the ratio of image tokens fed into the encoder. 'Reconstruct Ratio' denotes the ratio of reconstructed image tokens. 'Epochs' and 'FT Acc.' denote pre-training epochs and the top-1 accuracy of fine-tuning.

model. We follow the benchmark ViTDet [26, 27], the pre-trained backbone is adapted to FPN [28] in the Mask R-CNN framework [14]. The resolution of the input image, learning rate, and layer decay are respectively set as 1024×1024, 3e-4 and 0.8. The model is fine-tuned for 25 epochs with a total batch size of 64. As finetuning all methods is heavy and has the potential risk of non-optimal results, we take other results from original papers.

4.2. Comparison to State-of-the-arts

Classification. Tab. 1 shows the comparison of ImageNet finetuning results between our model and previous state-of-the-art approaches of the similar model size. We also list the Forward Ratio and Reconstruction Ratio in Tab. 1 to intuitively compare the role of encoder in each paradigm. For example, the encoders of BEiT V1/V2 [3, 33], MaskFeat [6] and MVP [44] process the full-size input with 40% masked token. During the encoding, these 40% masked tokens will be filled under the supervision of reconstruction targets. This kind of paradigm includes invalid information in inputs, leading to efficiency damage and risks of a gap between pre-training downstream tasks. For MAE [13] and MILAN [17], the encoder and decoder respectively process 25% and 100% patches. This paradigm still suffers from the computation on the decoder as it is totally abandoned in downstream tasks. A recent work FD-CLIP [45] uses the normalized feature of CLIP for distillation, thus the encoder process the full-size input. Compared to them, our model only processes 30% patches. While significantly reducing the training cost (1/3 Forward Ratio and 2/3 PT Epochs), our model outperforms BEiT V2, MaskDistill and FD-CLIP on Top-1 Accuracy (85.4% vs. 85.0% and 84.9%).

Detection and Segmentation. To verify the generalization of our method, we evaluate on COCO object detection and instance segmentation, by adapting the pre-trained ViT-B/16 backbone to FPN [28] in the Mask R-CNN framework [14]. The results are shown in Tab. 2. Our model achieves 52.1% on AP^{box} and 45.7% on AP^{mask}. Our model outperforms ViTDet [26] with shortened pre-training epochs from 1600 to 400 and fine-tuning epochs from 100 to 25.

4.3. Ablation Study

Dynamic Alignment. To verify the effectiveness of the Dynamic Alignment (DA) module, we conduct a series of experiments of making comparisons between w/ and w/o DA and different top Ks. The results are shown in Tab. 3. To fairly compare, We pre-train all models on ImageNet for 200 epochs, under the setting of 70% mask ratio with Attentive Masking strategy and CLIP-B/16 as the teacher model. For alignment type, Dynamic denotes using our proposed DA module and Layer-wise denotes aligning features layer-by-layer without a Dynamic Alignment Matrix. From Tab. 3, Multi-level alignment target performs better than only aligning with the teacher model's final output (85.3% vs. 84.9%). Multi-level features from the teacher model can provide richer semantics as supervision. It's intuitive that the latter blocks of the teacher model provide more high-level than former ones. As a result, a proper k has a big impact. We search the best of 3, 5, 7 for the hyper-

Method	Supervision	PT Epochs	AP^{box}	AP^{mask}
Supervised [14]	IN-1K Label	-	47.9	42.9
MoCov3 [6]	-	300	47.9	42.7
DINO [4]	-	300	46.8	41.5
BEiT [3]	DALLE	300	42.6	38.8
PeCo [7]	dVAE	300	43.9	39.8
SplitMask [9]	dVAE	300	46.8	42.1
CAE [5]	DALLE	800	49.2	43.3
MAE [13]	RGB	1600	50.3	44.9
ViTDet [26]	RGB	1600	51.2	45.5
MILAN [17]	CLIP-B	400	**52.6**	45.5
Ours	CLIP-B	400	52.1	**45.7**

Table 2. **Object detection and instance segmentation results** obtained by finetuning Mask R-CNN [14] on MS-COCO dataset [29]. All numbers are reported from the original paper and our model is implemented with the ViTDet protocol. All methods use ViT-B/16 pre-trained on ImageNet-1K dataset as the backbone. Our model outperforms ViTDet [26] with shorten pre-training epochs from 1600 to 400. "PT Epochs" refer to the pre-training epochs.

Model	Align Type	Top K	FT Acc.(%)
ViT-B/16	Dynamic	1	84.9
		3	85.1
		5	**85.3**
		7	85.1
	Layer-wise	1	84.8
		5	85.1

Table 3. Ablation study of alignment strategy. MaskAlign achieves best results by Dynamic Alignment on Top 5 level feature. We pre-train all models on ImageNet for 200 epochs, under the setting of 70% mask ratio with Attentive Masking strategy.

parameter top K, and keep it fixed for other experiments in this paper. Compared with Layer-wise type, Dynamic type achieves better performance (85.3% vs. 85.1%). This validates the effectiveness of our Dynamic Alignment (DA) module. It's worth noting that Layer-wise type with top 1 is equivalent to feature distillation on visible tokens, and our method gains 0.5% improvement on it. These results verify that the combination of MaskAlign and Dynamic Alignment will significantly improve the pre-training.

Mask Strategy. As our paradigm is totally different from existing paradigms based on reconstruction, the impacts of mask ratio could be also different. In this part, we compare different mask ratios and strategies. As MaskAlign only uses $1 - r\%$ features produced by the teacher model per iteration, different mask ratios will result in unmatched training efficiency. To make a fair comparison, we adjust the training iterations for each experiment to be equal to 200 epochs for 70% mask ratio. We pre-train ViT-B/16 as the student model under the setting of Top 5 Dynamic Alignment with CLIP-B/16 as the teacher model. The results are

Model	Mask Type	Mask Ratio	FT Acc.(%)
ViT-B/16	Attentive	0%	84.9
		20%	85.1
		40%	85.2
		60%	**85.3**
		70%	**85.3**
		75%	85.1
		80%	85.1
		90%	84.4
	random	70%	85.1

Table 4. Ablation study of mask ratio and strategy. MaskAlign performs well on wide range of mask ratios. We pre-train all models under the setting of Top 5 Dynamic Alignment with CLIP-B/16 teacher model.

shown in Tab. 4. From Tab. 4, interestingly, we observe that although the peak accuracy is gained at around 60-70%, the performance and mask ratio curve is much more flat than MAE's [13]. For example, when mask ratio change from 60% to 20%, the FT Acc. of MAE drops 1.6% while ours only drops 0.2%. And the curve is also flat at high mask ratios. For reconstruction, excessive visible patches will lead to a simplistic pre-training task while insufficient visible patches will fail in providing necessary information for the model to infer masked patches. As a result, reconstruction-based methods need more careful choice of mask ratios. For alignment on the visible features, even a 0% mask ratio gains improvement compared with the teacher model [45]. Introducing masked modeling into distillation will further improve the performance, yet relies less on mask ratio.

Teacher Model. To study the ability of generalization and scaling-up behavior, we also conduct experiments on com-

Model	Teacher Model	T-FT Acc.	S-FT Acc.
ViT-B/16	DINO-B	82.8 [4]	83.9 (+1.1)
	CLIP-B	82.9 [45]	85.3 (+2.4)
	CLIP-L$_{196}$	-	85.6
ViT-L/16	CLIP-B	82.9 [45]	86.5 (+3.6)
	CLIP-L$_{196}$	-	87.4

Table 5. Comparison of different teacher model. T-FT and S-FT denote the finetuning Acc. of teacher model and student model. CLIP-L$_{196}$ denotes input 196×196 resolution image to CLIP-L. We pre-train all models on ImageNet for 200 epochs.

paring impacts by different teacher models. In this part, we pre-train each model for 200 epochs, under the setting of Top 5 Dynamic Alignment and 70% mask ratio. We choose DINO-B/16 [4], CLIP-B/16 and CLIP-L/14 as teacher models to train ViT-B/16. To match the student model's feature size, we resize the input resolution of CLIP-L/14 to 196×196. To study the ability of generalization on larger models, we choose CLIP-L/14 as the teacher model to train a ViT-L/16 model. The results are shown in Tab. 5. We find that MaskAlign consistently works well on various teacher models. Student models outperform teacher models in the same size: for base size models, MaskAlign leads to 1.1% and 2.4% improvement on DINO and CLIP, respectively. For a large-size model ViT-L/16, MaskAlign leads to 3.6% improvement on CLIP-B/16. These results validate that our method has a good ability to generalize on different teacher models.

Adaptation Details. We also compare different Adaptors and target normalization methods. For Adaptor, we compare a simple Linear layer with a 2-layer MLP, with the same hidden size as the dimension of features from the student model. For normalization, we compare LayerNorm, BatchNorm and original feature. We disable learnable scale and bias for both LayerNorm and BatchNorm. We pre-train all models on ImageNet-1K for 200 epochs, under the setting of 70% mask ratio with Attentive Masking strategy. Top 5 Dynamic Alignment with CLIP-B/16 teacher model. From the results in Tab. 6, MLP adaptor does not bring performance improvement compared to a simple Linear Layer, even though more parameters are included in alignment. Our results show that the type of normalization has a big impact, LayerNorm performs much better than BatchNorm and no normalization. This observation is also consistent with [45]. We find that including [CLS] token in alignment will slightly improve the accuracy. As [CLS] of CLIP aggregates global information of the image, mimicking [CLS] helps MaskAlign learn more knowledge of the interaction between different patches. Results in Tab. 3, 4 and 5 are under w/o [CLS] token setting.

Model	Adaptor	Norm	[CLS]	FT Acc.(%)
ViT-B/16	Linear	LN	w/	**85.4**
	Linear	LN	w/o	85.3
	MLP	LN	w/o	85.2
	Linear	BN	w/o	84.7
	Linear	w/o	w/o	84.3

Table 6. Ablation study of other adaptation details. We pre-train all models on ImageNet for 200 epochs, under the setting of 70% mask ratio with Attentive Masking strategy. Top 5 Dynamic Alignment with CLIP-B/16 teacher model.

Model	Epochs	Time	FT Acc.(%)
MVP [44]	300	2.8×	84.4
BEiT V2 [33]	300	2.8×	85.0
MILAN [17]	400	3.2×	**85.4**
FD-CLIP [45]	300	2.8×	84.9
Ours	200	1.0×	**85.4**

Table 7. Comparison of pre-training speed. MaskAlign achieves SOTA results with only about 1/3 training time. "Epochs" refer to the pre-training epochs of various methods. We compare with existing works with CLIP-ViT-B/16 as teacher model. Training speeds of these works are uniformed.

Pre-training Speed. We compare the pre-training speed of our model with state-of-the-art methods. MVP [44], BEiT V2 [33] and MILAN [17] are all Masked Image Modeling pre-training models using CLIP-ViT-B/16 features as reconstruction target. Specifically, MVP and BEiT V2 are in inpainting-style. They mask input patches by replacing with learnable tokens then filling these patches, thus the encoder processes a full-size input. MILAN adopts decoder-style, the encoder only processes visible patches and full-size intermediate features are handled by a decoder. Although the decoder is much lighter than the encoder, its full-size input still brings much computation. Besides, the decoder only plays a role in pre-training stage, and is abandoned in downstream tasks. This gap limits the efficiency of the pre-training. A recent work FD-CLIP [45] distills the feature map of CLIP model at the output. Compared with them, our model has the lightest architecture in theory.

Tab. 7 lists the pre-training times and ImageNet-1K finetuning performances. The training times are uniformed. Our model outperforms MVP, BEiT V2 and FD-CLIP, and achieves comparable performances with MILAN, but with only about 1/3 training time. It's worth noting that our model only uses $1 - r\%$ features produced by the teacher model per iteration. As a result, the inference time of the teacher model makes our method not have a linear acceleration ratio. Our method's acceleration is more significant in scenarios when teacher model is lighter than the student

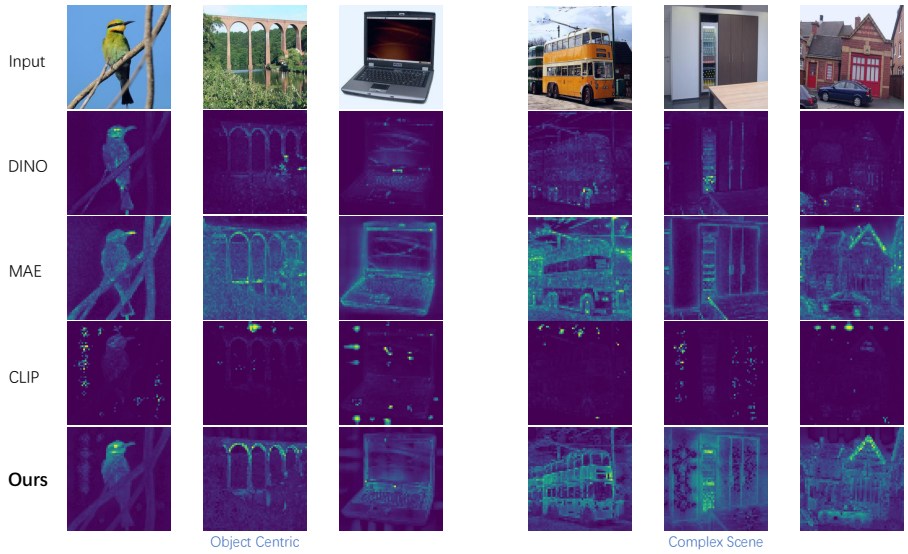

Figure 3. Visualization of attention weights at the last self-attention layer in DINO [4], MAE [13], CLIP [35] and ours. Object centric images are on the left while complex scenes are on the right. Our model highlights more reasonable regions.

model, e.g., CLIP-B/16 supervises ViT-L/16.

Attention Visualization. To intuitively peek at what is learned in MaskAlign during pre-training, we visualize the attention map of [CLS] token of the last self-attention layer of different models. The comparison is shown in Fig. 3. DINO is trained by contrastive learning, which minimizes the similarities between augmented views of images. Random cropping in view augmentation makes DINO tend to focus mainly on the salient region in the original image. Thus the attention weights of DINO are usually concentrated on one salient object. MAE reconstructs masked pixels from visible regions, thus more texture information is learned, leading to a waste of capacity on low-level features irrelevant for semantic understanding. CLIP has good semantic alignment with language, however, we surprisingly find that CLIP features have bad correspondence to semantic regions. This may caused by its sparse supervision of texts. Although supervised by CLIP, our model's attention map seems more reasonable on both object-centric images and complex scenes. MaskAlign accurately concentrates on salient objects. While handling complex scenes, MaskAlign covers different semantic regions in one image.

5. Conclusion

In this paper we first categorize and rethink existing Masked Image Modeling (MIM) paradigms. Both inpainting-style and decoder-style models need much computation on masked tokens, decreasing the training efficiency of pre-training. Following some approaches that apply semantic-rich teacher models to extract image features as supervision, we propose a MIM paradigm named MaskAlign without any reconstruction. MaskAlign simply aligns the visible features extracted by the student model and intact image features extracted by the teacher model. And we propose a Dynamic Alignment (DA) module to tackle the problem of input inconsistency between the student and teacher model. We conduct extensive experiments to verify the effectiveness of our method. Our model achieves state-of-the-art performances with much higher pre-training efficiency. In the future, we will explore the scaling up of MaskAlign for vision recognition.

6. Broader Impact

Our work explores a new paradigm for masked image modeling, which may encourage future works to reconsider the role of masking in pre-training or distillation. On contrary of existing models based on reconstruction, MaskAlign borrows the teacher model's semantic information to learn feature consistency and without any supervision on masked tokens, our method demonstrates a strong ability of representation and efficiency. Besides, MaskAlign has an extremely light and simple framework. In the future, we may move forward to 1) find more mathematical explanations of MaskAlign, and 2) transfer MaskAlign to large-scale multi-modal pre-training to leverage its advantages in both efficiency and simplicity

Acknowledgement This work is partially supported by the National Natural Science Foundation of China (Grant No.62206272), National Key R&D Program of China (NO.2022ZD0160100), and in part by Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

References

[1] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 5

[2] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. *arXiv preprint arXiv:2208.12256*, 2022. 5

[3] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2021. 1, 2, 4, 5, 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1, 3, 6, 7, 8

[5] 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. *arXiv preprint arXiv:2202.03026*, 2022. 1, 4, 5, 6

[6] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 1, 2, 3, 4, 5, 6

[7] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 2, 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1, 3

[9] Alaaeldin El-Nouby, Gautier Izacard, Hugo Touvron, Ivan Laptev, Hervé Jegou, and Edouard Grave. Are large-scale datasets necessary for self-supervised pre-training? *arXiv preprint arXiv:2112.10740*, 2021. 6

[10] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[11] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Qiao. Convmae: Masked convolution meets masked autoencoders. In *NeurIPS*, 2022. 1, 4, 5

[12] Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention. *arXiv preprint arXiv:2303.02995*, 2023. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 8

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5, 6

[15] Tong He, Chunhua Shen, Zhi Tian, Dong Gong, Changming Sun, and Youliang Yan. Knowledge adaptation for efficient semantic segmentation. In *CVPR*. 3

[16] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 3

[17] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 2, 3, 5, 6, 7

[18] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2

[19] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. 4

[20] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 5

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2

[22] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, 2020. 3

[23] Ioannis Kakogeorgiou, Spyros Gidaris, Bill Psomas, Yannis Avrithis, Andrei Bursuc, Konstantinos Karantzalos, and Nikos Komodakis. What to hide from your students: Attention-guided masked image modeling. *arXiv preprint arXiv:2203.12719*, 2022. 3

[24] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1, 2

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[26] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 5, 6

[27] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 5

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[29] 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, 2014. 4, 6

[30] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022. 5

[31] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2

[32] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, pages 2085–2094, 2021. 2

[33] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1, 2, 3, 4, 5, 7

[34] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. A unified view of masked image modeling. *arXiv preprint arXiv:2210.10615*, 2022. 2, 5

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3, 8

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831, 2021. 2

[37] Jason Tyler Rolfe. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200*, 2016. 2

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 4, 5

[39] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. *arXiv preprint arXiv:2210.06031*, 2022. 2

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. 1, 3

[41] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, pages 32–42, 2021. 1

[42] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 3

[43] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 1, 5

[44] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *ECCV*, 2022. 1, 2, 3, 4, 5, 7

[45] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 2, 3, 4, 5, 6, 7

[46] 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 *CVPR*, 2022. 1, 2, 5

[47] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5036–5045, 2022. 2

[48] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. In *NeurIPS*, 2021. 3

[49] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 2

[50] Jing Yang, Brais Martinez, Adrian Bulat, Georgios Tzimiropoulos, et al. Knowledge distillation via softmax regression representation learning. ICLR, 2021. 3

[51] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 2, 3

[52] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *ICML*, 2021. 2