DPPMask: Masked Image Modeling with Determinantal Point Processes

Junde Xu [1,2,3*] Zikai Lin[1,2*] Donghao Zhou[1,2] Yaodong Yang[4] Xiangyun Liao[1,2]
Qiong Wang[1,2] Bian Wu[3] Guangyong Chen[3†] Pheng-Ann Heng[4]
[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Zhejiang Lab [4]The Chinese University of Hong Kong

Abstract

Masked Image Modeling (MIM) has achieved impressive representative performance with the aim of reconstructing randomly masked images. Despite the empirical success, most previous works have neglected the important fact that it is unreasonable to force the model to reconstruct something beyond recovery, such as those masked objects. In this work, we show that uniformly random masking widely used in previous works unavoidably loses some key objects and changes original semantic information, resulting in a misalignment problem and hurting the representative learning eventually. To address this issue, we augment MIM with a new masking strategy namely the DPPMask by substituting the random process with Determinantal Point Process (DPPs) to reduce the semantic change of the image after masking. Our method is simple yet effective and requires no extra learnable parameters when implemented within various frameworks. In particular, we evaluate our method on two representative MIM frameworks, MAE and iBOT. We show that DPPMask surpassed random sampling under both lower and higher masking ratios, indicating that DPPMask makes the reconstruction task more reasonable. We further test our method on the background challenge and multi-class classification tasks, showing that our method is more robust at various tasks.

1. Introduction

Self-supervised learning aims to extract semantic features by solving auxiliary prediction tasks (or pretext tasks) with pseudo labels generated solely based on input features. While various tasks have been proposed for self-supervised learning, one intuitive idea is learning representations by recovering the original data from the corrupted structure. The philosophy behind such methods is simple: what the model generates can examine whether the model understands. This principle was first introduced by the denoising autoencoder [57] which has supported significant advances in NLP [4, 37, 48]. Methods that follow this idea such as BERT [15] now have become a dominant routine. In the field of image processing tasks, although reconstruction-based pre-training was first put forth by [45], it wasn't until recently that methods based on this concept were brought back to state-of-the-art performance. Benefiting from the new network architectures like ViT [17], Masked Image Modeling (MIM) has become highly popular, and there is a series of more aggressive masking strategies like MAE [26], simMIM [62].

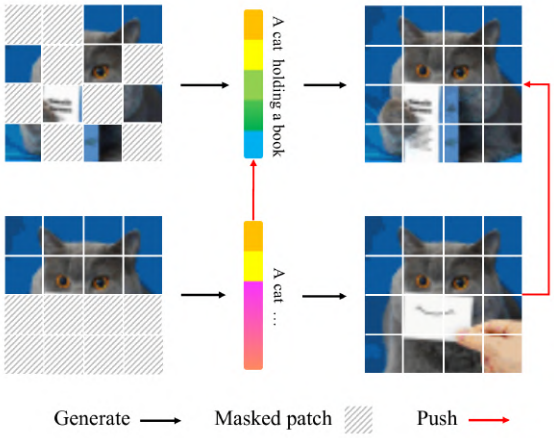

Figure 1. Illustration of the misalignment problem in MIM. The model generates predictions that differ plausibly from the original image, while the original image has still been imposed as supervision, leading to an unreasonable high loss.

However, simply random masking can be problematic in practice. An important fact is that it is unreasonable to force the model to reconstruct something beyond recovery. Consider a simple case as Fig. 1. The top row of Fig. 1 shows the underlying logic of MIM: successful reconstruction implies the network captures the correct semantic features. While

*Equal contribution
†Corresponding Author (gychen@zhejianglab.com).

the bottom row of Fig. 1 shows a failure case: if the masking process happens to drop an important semantic of the original image, the book in Fig. 1, then will change the semantics of the original image and make the network hardly recover it from the rest. In this case, if the model continues to be forced to reconstruct the original image, the model might fill in the obscured image with whatever is feasible, which will interfere with the process of learning the original features. Furthermore, as the masking rate increases, the original semantic information is distorted with a higher probability. We refer to such a situation as a misalignment problem, *i.e.* the semantics of masked image and the original image are miss-aligned. Consequently, the misalignment problem will cause the alignment of improper sample pairs, which will eventually harm the performance of the downstream task.

Some studies also proposed constructing better sample pairs for MIM. In MaskFeat [59], they change the pixel reconstruction task to HOG reconstruction, to reduce the impact of some ambiguous situations for the network to prediction, such as colors, and textures. However, MaskFeat can only reduce the impact of hardly recoverable high-frequency signals. If the masked part contains the whole object instance, there is still not enough information for the network to rebuild. In ADIOS [52] and SemMAE [35], they train an extra segmentation network to partition the image into different semantics. However, the number of semantics varies significantly in different images, thus, it is hard to find an optimal semantic partition network. In AttMask [29], they mask the most attended patches according to their attention score to construct more challenging MIM tasks. However, AttMask needs an attention map to perform sampling, which can not fit into reconstruction-based MIM methods (e.g., MAE) seamlessly. Recently, AMT [25] adding attention map guided masking to MAE. Unfortunately, these algorithms do not take into account the misalignment problem, which leads to inferior performance.

In contrastive learning, an *InfoMin* principle suggests that two augmented views of an image should retain task-relevant information while minimizing irrelevant nuisances [54]. Analog to MIM, we can summarize the following two conditions: First, the selected patches should be representative enough to cover the whole semantic information of the original image. Second, the masking ratio should be set to a high level to minimize the irrelevant information shares between different masks of the same image. While the second constraint is easy to satisfy, the problem is how to retain the task-relevant original semantics under the limited input ratio. To address this, we propose a novel masking strategy based on Determinantal Point Process (DPPs). DPPs are elegant probabilistic models on sets that can capture both quality and diversity when a subset is sampled from a ground set [32, 33], making them ideal for modeling the set that contains more information of original images as possible. During the sampling process, DPPs will compute the distance of each patch, and select patches that are dissimilar from the selected subset. This process makes the network focus on the patches with more representative information. For example, the unique color, texture, etc. We show that our new sampling strategy can obtain more representative patches to keep the semantics unchanged and alleviate the impact of the misalignment problem. More importantly, We show that DPPMask surpassed random sampling under both lower and higher masking ratios, indicating that DPPMask makes the reconstruction task more reasonable. Furthermore, our method needs no extra training process and achieves minimal computational resource consumption.

Our contribution can be summarized as follows:

- We analyze the training behavior of reconstruction-based MIM and discuss the impact of the misalignment problem.
- To alleviate the impact of misalignment in MIM, we proposed a novel plug-and-play sampling strategy called DPPMask based on DPPs. Our method can generate more reasonable training pairs, is simple yet effective, and requires no extra learning parameters.
- We verify our method on two representative MIM frameworks, our experiments evidence that features learned by fewer misalignment problems achieve better performance in downstream tasks.

2. Related Work

Self-supervised learning. Classic deep learning trains the parameters of the model by utilizing labeled data. Instead, self-supervised learning(SSL) expects to acquire representations with unlabeled data by a pre-text task. Among them, Masked language modeling (MLM) has taken the lead to be a highly influential self-supervised learning model before. $e.g.$,, BERT [15] and GPT [47, 48] are such successful methods that the academia has focused on these two models for pre-training in NLP. These models leverage visible tokens in a sequence and predict invisible tokens to gain appropriate representations, which have been proved to successfully repaint the field [4]. In other fields of SSL, there have been numerous methods that focus on different pretext tasks like reconstructing original tokens from image/patch operations [5, 16, 21, 43, 60]and Spatio-temporal operation [19, 23, 41, 44, 58]. A well-known method is contrastive learning that capitalizes on augmentation invariance in the feature space and could be evaluated by linear probing [6, 7, 11, 12, 18, 27, 46], which was the previous mainstream based on SSL.

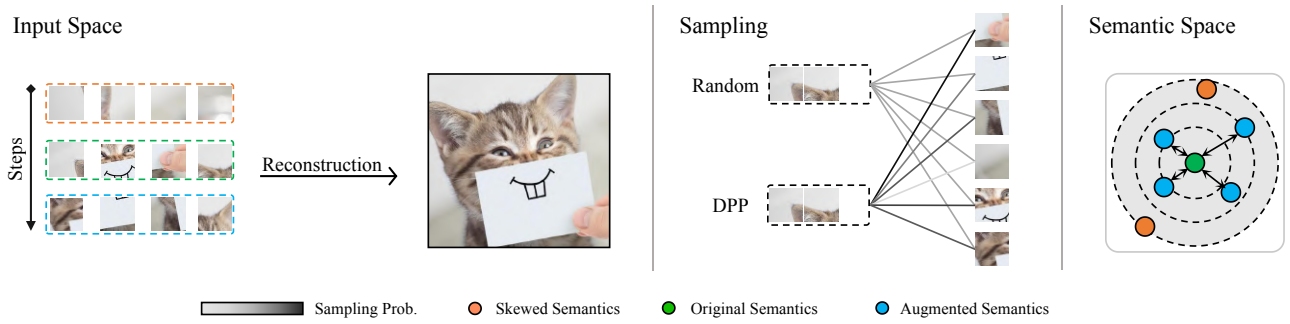

Figure 2. An illustration of misalignment and our method. For each image, the masking policy can propose different patch sets for reconstruction, however, some of them with skewed semantics are not suitable for reconstruction. Our method models the probability of co-occurrence of each patch to avoid such misalignment problems.

Masked Image Modeling(MIM) Masked Image Modeling recently has shown capability to reconstruct pixels [1] from corrupted images. MIM can be seen as a generalized Denoising AutoEncoders (DAE) [9, 56], which aims to reconstruct the original tokens from corrupt input. e.g., inputing missing color channels [66] or missing pixels [57]. Context Encoder [45] reconstructs a rectangle area of the original images using convolutional networks. Then ViT [17] and iGPT [10] recall the learning approach of predicting patches with a contrastive predictive coding loss on the modern vision Transformers, and show strong potential in representation learning. BEiT [2] proposes to use a pre-trained discrete VAE [49]as the tokenizer, and improves MIM's performance further. However, the tokenizer needs to be offline pre-trained with matched model and dataset which limits its adaptivity. To end this, iBOT [67] presents a new framework that performs masked prediction with an online tokenizer and gains prominence achievement. Recently, equipped with a more aggressive masking strategy, SimMIM [62] and MAE [26] further demonstrate that simple pixel reconstruction can achieve competitive results from previous pre-training methods.

Determinantal Point Process. Determinantal point processes (DPPs) are probabilistic models of configurations that favor diversity [40]. Its repulsion brings new potential to enhance diversity in multiple machine learning problems, such as feature extract from high dimensional data [3], texture synthesis in image processing [34], building informative summaries by selecting diverse sentences [32].

In addition to DPPs, there are some previous methods like Markov random fields (MRFs). However, MRF assumes that repulsion does not depend on context too much, so it cannot express that, say, there can be only a certain number of selected items overall [32]. The DPPs can naturally implement this kind of restriction through the rank of the kernel.

3. Methods

In this section, we first discuss the impact of the misalignment problem on the downstream tasks, then, we give a brief introduction to DPPs. Finally, we propose our method of applying DPPs in the patch masking process.

3.1. Misalignment in MIM

Let x_1, x_2, z be the masked image, reconstruction target, and hidden vector. In the ideal situation, the input x_1 and reconstruction target x_2 all followed an identical distribution $x_1, x_2 \sim P(z)$. However, consider the input only captures the partial information of the original image, in this case, $x_1 \sim P(z')$ where z' defines a different distribution to z. For pixel reconstruction tasks, we denote the encoder and the decoder parameterized by ϕ and θ respectively. Following [2], MIM training can be viewed as variational autoencoder training [30], which can be described as a Maximum A Posteriori (MAP) Estimation:

$$\arg\max_{\phi,\theta} \mathbb{E}_{z' \sim q_\phi(z'|x_1)} \log p_\theta(x_2|z') \tag{1}$$

Suppose the encoder is capable of capturing the semantic information of x_1 and the decoder is capable of recovering the image described by z and z', by unfolding the decoder part, we get:

$$p_\theta(x_2|z') = \int p_\theta(x_2|z) P(z|z') \mathrm{d}z \tag{2}$$

The Eq.1 indicates that the pixel reconstruction task is minimizing the distance between representations of masked image z' and original image z. Now, let us zoom the lens to multiple steps, MIM can receive multiple different masked images of the original image. By training the network to reconstruct original images from different masked ones, MIM minimizes the distance between those different masks. Fig. 2 shows a diagram of such training behavior, where different masking result lies at a different location of

a hyperplane and construct the semantic space. During the reconstruction training process, data points in the semantic space are pushed to align with the location of the original image. Consequently, as they are aligned with the same data point, the distance between each data point in semantic space is also minimized. While misalignment is a false aggregation of data points that have skewed semantics (orange dot in Fig. 2). For other MIM reconstruction targets, such as visual tokens in BEiT [2] and iBOT [67], it is easy to verify that they also share similar training behavior. A similar conclusion has also been reported in [65], however, they focus on the dimensional collapse issue in MAE and neglect the misalignment problem of MIM. In practice, semantics are not evenly distributed in images. These semantics are likely to be ignored by random masking strategies. As the MIM pulls masked samples together, two images with different semantics are miss-aligned. If the changed semantics is an important clue for image understanding, such a problem can seriously affect downstream performance. To this end, we propose a new sampling strategy to select as representative patches as possible.

3.2. Determinantal Point Process

Our core technical innovation is modeling the patch masking process with DPPs. To this end, we start with a high-level overview of DPPs.

Brief intro. A determinantal point process (DPPs) is a distribution over configurations of points. The defining characteristic of the DPP is that it is repulsive, which makes it useful for modeling diversity [32]. Formally, a point process $\mathcal{P}$ on a discrete set $S = \{1, 2, ..., N\}$ is a probability measure on 2^S, the set of all subsets of S. $\mathcal{P}$ is called a determinantal point process if, when A is a random subset drawn according to $\mathcal{P}$, we have,

$$\mathcal{P}(Y = A) \propto \det(L_A), \tag{3}$$

where $L \in R^{N \times N}$ is a real, symmetric, positive semi-definite kernel, and $L_A \in R^{|A| \times |A|}$ is a submatrix of L indexed by elements of A. Note this is an unnormalized probability of sampling a set of A. The normalization constant is defined as the sum of the unnormalized probabilities over all subsets of the S, *i.e.* $\sum_{A \subseteq S} \det(L_A)$. We can compute the normalized constant by the following theorem [32]:

Theorem 1 *For any $A \subseteq S$:*

$$\sum_{A \subseteq Y \subseteq S} \det(L_Y) = \det(L + I_{\bar{A}}), \tag{4}$$

where $I_{\bar{A}}$ is a diagonal matrix such that $I_{ii} = 0$ for indices $i \in A$ and $I_{ii} = 1$ for $i \in \bar{A}$.

Setting $A = \emptyset$, we obtain the following corollary:

(a) $\tau = 0$ (b) $\tau = 0.9$ (c) $\tau = 1$

Figure 3. Qualitative comparison of different purge ratios, a suitable purge ratio can preserve the semantics of original images while maintaining the variety of augmented inputs.

Corollary 1.1

$$\sum_{A \subseteq S} \det(L_A) = \det(L + I_S). \tag{5}$$

Therefore, for any $A \subseteq S$, we can compute its probability by:

$$\mathcal{P}(Y = A) = \frac{\det(L_A)}{\det(L + I)}, \tag{6}$$

where I is the identity matrix.

We give a simple example to illustrate how DPPs model diversity. Suppose we have two patches i, j to select, *i.e.* $A = \{i, j\}$, we denote the vector of each patch as $S_i, S_j \in R^{1 \times n}$, where n is the dimension of elements in S. We can compute the L-ensemble $L_{ij} = S_i^T S_j$ and their co-occurrence probability by Eq.6. The numerator of Eq.6 can be written as: $det(L_A) = L_{ii} \times L_{jj} - L_{ij} \times L_{ji}$. Note that L_{ij} and L_{ji} measure the similarity between elements i and j, being more similar lowers the probability of co-occurrence. On the other hand, when the subset is very diverse, *i.e.* $L_{ij} \times L_{ji}$ becomes small, the determinant is bigger and correspondingly its co-occurrence is more likely. The DPP thus naturally diversifies the selection of subsets.

Unfortunately, to our best knowledge, the implementation of exact DPP sampling needs matrix decomposition [20, 32], which is an unacceptable computation cost during the training iteration. Thus, to apply DPPs in MIM, an approximation is needed.

Greedy approximation. Considering we only select the subset Y with the highest probability under a cardinality constraint n, then such problem can be defined as:

$$Y_{\text{MAP}} = \arg\max_{Y \subseteq S} \det(L_Y). \tag{7}$$

This problem is known as maximum a posteriori (MAP) inference and has been proved as an NP-hard problem in DPPs [31]. Instead, the greedy algorithm is widely used for approximation [42] for MAP inference, justified by the fact that the log-probability of set in DPPs $f(Y) = \log det(L_Y)$ is sub-modular [22]. Thus, the selection process of our method can be described as follows:

$$j = \arg\max_{i \in S \setminus Y_g} f(Y_g \cup \{i\}) - f(Y_g), \tag{8}$$

where Y_g is the subset of Y. In each iteration, we add an

Algorithm 1 Greedy DPPs Sampling for MIM

Require: image patches S, Purge ratio τ, subset length N;
$S \leftarrow \text{shuffle}(S)$
$L \leftarrow \text{kernel}(S)$
$Y_g = []$
$d = \text{zeros}(\text{len}(S))$
while $N \geq 0$ **do**
  $d \leftarrow \text{update}(L, Y_g, d)$ ▷ Follow [8]
  if $\max(d) \geq \tau$ **then**
    $Y_g.\text{append}(\text{argmax}(d))$
  else
    $Y_g.\text{append}(\text{randomSelect}(d))$
  end if
  $N \leftarrow N - 1$
end while

item that maximizes the marginal gain to Y_g, until the maximal marginal gain emerges negative or goes against the cardinality constraint. We adopt a fast implementation of the greedy MAP inference algorithm for DPPs following [8]. Formally, since L is a PSD matrix,

the Cholesky decomposition of L_{Y_g} is available as

$$L_{Y_g} = VV^T, \tag{9}$$

where V is an invertible lower triangular matrix. For any $i \in Z \setminus Y_g$, the Cholesky decomposition of $L_{Y_g \cup \{i\}}$can be derived as:

$$L_{Y_g \cup \{i\}} = \begin{bmatrix} L_{Y_g} & L_{Y_g,i} \\ L_{i,Y_g} & L_{ii} \end{bmatrix} = \begin{bmatrix} V & 0 \\ c_i & d_i \end{bmatrix} \begin{bmatrix} V & 0 \\ c_i & d_i \end{bmatrix}^\top,$$

where row vector c_i and scalar $d_i \geq 0$ satisfies:

$$Vc_i^\top = L_{Y_g,i}, \qquad d_i^2 = L_{ii} - \|c_i\|_2^2 . \tag{10}$$

Then the determinate of $L_{Y_g \cup \{i\}}$ can be written as

$$\det\left(L_{Y_g \cup \{i\}}\right) = \det\left(VV^\top\right) \cdot d_i^2 = \det\left(L_{Y_g}\right) \cdot d_i^2. \tag{11}$$

Therefore, Eq.7 is equivalent to select the element i with maximum ${d_i}^2$.

After solving the equation, the Cholesky factor of L_{Y_g}can therefore be efficiently updated after a new item is added to Y_g. With these approximations, the selecting process can be fit in the GPU training loops. In our experiments, the acceleration ratio is up to 10 times faster with respect to exact DPPs sampling and brings the time cost of DPPs to the same level of random, more details can be found in the supplementary.

3.3. Purge misalignment with DPPs

In this section, we introduce two key factors of DPPMask: kernel and purge ratio.

Kernels. A common-used type of kernel is the class of Gaussian kernels [34, 55]. Defined by

$$\forall S_i, S_j \in S, \qquad L_{ij} = \exp(-\frac{||S_i - S_j||^2}{\epsilon}). \tag{12}$$

Where ϵ is called the bandwidth or scale parameter. This kernel depends on the squared Euclidean distance between the intensity values of pairs of patches. It is often used as a similarity measure on patches. The value of the parameter ϵ has an impact on how repulsive the DPPs are. However, note we only choose patches that maximize Eq.8, and the value of ϵ influences a little to model performance. This is because ϵ will *not* change the order of distances between patches. Thus, for better numerical stability, we normalize each patch before computing the distances and set ϵ to 1 empirically.

Purge ratio. As shown in Fig. 2, DPPMask aims to purge those cases in that semantic information has been changed by masking. However, DPPMask can get over-purged in some cases. As Fig. 3 shows, due to patches of the sky being too similar to each other, then the greedy selection will only focus on the foreground, as it is more diverse than the background. This situation makes the MIM task too easy and purges most of useful augmented inputs, which is not helpful in feature learning. A simple modification can tackle this problem. Instead of letting the selection process hit the cardinality constraint, we set a parameter called purge ratio $\tau \in (0, 1)$ as the threshold of maximal marginal gain. Concretely, in each iteration, we monitor the distance of the next patch to the selected subsets, if the distance is below the purge ratio τ, abort the greedy selection process and fill the subset will random patches. The purge ratio plays a role of adjust how "severe" the DPPMask is. In particular, $\tau = 0$ indicates the selection process becomes fully greedy and $\tau = 1$ indicates fully random sampling. Fig. 3 shows greedy selection under three different purge ratios, a higher purge ratio can prevent DPPMask get over-purged and maintain the input diversity for training the network.

In this section, we first analyze the training behavior of MIM, then we give our method namely DPPMask which uses DPPs to model the repulsion of image patches, in order to sample the most representative patches and preserve the original semantic information of images. We summarize our algorithm in Alg. 1.

4. Experiment

4.1. Implementation details

To examine the effectiveness of our method, we perform DPPMask on two representative MIM methods: MAE [26],

Method	Pre-train loss	Linear prob.	Fine-tuned
$\tau = 0.6$	0.417	-	89.67
$\tau = 0.8$	0.434	62.58	**89.67**
$\tau = 0.9$	0.440	63.22	89.56
MAE	0.444	**67.08**	89.45

Table 1. Detailed results of MAE+DPPMask on ImageNet-100. The best of each metric are marked in bold.

Method	NMI	ACC	Linear prob.	Fine-tune
$\tau = 0.6$	0.518	67.88	72.84	87.58
$\tau = 0.8$	0.522	68.04	73.56	87.64
$\tau = 0.9$	**0.525**	**68.68**	**73.60**	**87.84**
iBOT	0.522	68.28	73.30	87.44
iBOT+AttMask	0.512	67.32	72.30	87.44

Table 2. Detailed results of iBOT+DPPMask on ImageNet-100. The best of each metric are marked in bold.

iBOT [67], which represent two different MIM frameworks: pixel reconstruction and feature contrast. For MAE, images are patched by convolutional kernels and added with a position embedding, after that, we compute the distances of patches. For iBOT, images are fed into a teacher model to get semantic tokens, which we compute distances based on. Compare to direct computing with pixel intensities, semantic tokens may contain more useful information for partitioning images. For example, to identify instances that share similar appearances.

We adopt two different scales of backbones, ViT-Base and ViT-Small for MAE and iBOT respectively. We mainly evaluate our algorithms on the ImageNet-100 dataset, which is derived from ImageNet-1K [51, 52]. We train MAE and iBOT for 400 and 100 epochs respectively. Unfortunately, our computational resources can not support us to make out larger-scale experiments, such as more training epochs and heavier backbones. We leave this for future work. We set the masking ratio is set to 0.75 for MAE and 0.7 with 0.05 variance for iBOT by default.

4.2. A detailed study of misalignment

We give our main result in Tab. 1 and Tab. 2. The final feature vector for classification is obtained by global pooling. For fine-tuning tasks, we run each setting with three random seeds and report their average performance. We train the iBOT model under the fine-tuning and linear probing parameter setting of MAE for reducing the experiment's complexity.

The performance gain brings by DPPMask. We first observed a steady performance gain on fine-tuning tasks in both MAE and iBOT frameworks. In MAE, we make 0.2% accuracy gains. In iBOT, we make 0.4% accuracy gains. This shows our method can improve the representation power by purging the misalignment samples. As the threshold τ increases, the pre-train loss becomes smaller in response. However, when τ is too low, greedy sampling goes over-purged and makes the task too simple for the network to learn useful features. Besides, iBOT and MAE show different preferences of τ, this is because we apply DPPs on the output of the teacher model, which is a different distribution from MAE. For linear probing, we notice a significant performance drop of MAE. This does not surprise us as we analyze in Sec. 3.1. DPPs make the sample space shrunk in order to purge the misalignment samples. Aggregating fewer samples together makes the feature space more continuous and less linear separable. Notably, the features also become more precise to describe an image, which can reflect the fine-tuned performance. This behavior is not observed on iBOT, which uses extensive augmentation to further expand the scale of positive samples. Instead, iBOT got 0.3% performance gain on linear probing as well as cluster performance (NMI and ACC), which proves our method indeed purged improper samples that hurt feature learning. Tab. 3 shows our method alongside with other advanced sampling methods on ImageNet-1K, our methods surpassed other sampling methods.

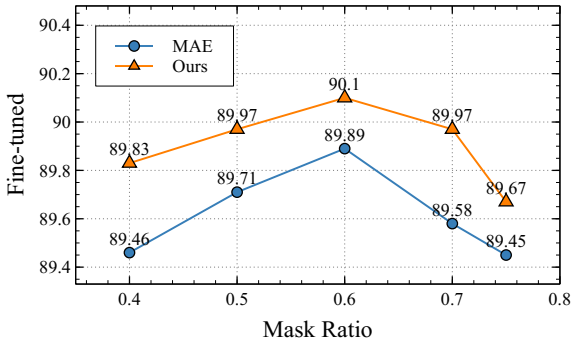

Figure 4. Validation accuracy of MAE on ImageNet-100 under different masking ratios.

Method	Backbone	Epochs	Fine-tuned
MAE [65]	ViT-Base	400	82.9
MAE+AMT [25]	ViT-Base	400	82.8
SemMAE [35]	ViT-Base	800	83.3
MAE+DPPMask	ViT-Base	400	**83.3**

Table 3. Comparison with other sampling methods on ImageNet-1K. The best of each metric are marked in bold.

DPPMask makes the MIM task more reasonable. Another key factor of successfully applying MIM is the masking ratio of input images [26, 62]. It should be high enough to construct a meaningful reconstruction target while preventing the task from degenerating to simply copying-pasting from neighboring patches. However, the root cause of the misalignment problem is also the aggressive mask-

(a) Orig. (b) O.F. (c) M.S. (d) M.R. (e) M.N.

Figure 5. Examples of background challenge.

Method	Orig.	O.F.	M.S.	M.R.	M.N
iBOT	56.05	42.32	48.07	39.70	37.23
iBOT+AttMask	54.22	41.01	46.15	37.98	36.02
iBOT+DPPMask	**56.47**	**43.36**	**49.65**	**40.59**	**38.44**
MAE	55.46	43.04	47.36	40.12	38.30
MAE+DPPMask	**56.32**	**44.12**	**47.73**	**40.74**	**39.43**

Table 4. Performance on background challenge.

Method	mAP	$F1_{all}$	$F1_{class}$
iBOT	63.16	64.94	55.78
iBOT+AttMask	63.26	65.31	56.36
iBOT+DPPMask	**63.78**	**65.72**	**56.40**
MAE	68.95	69.05	61.24
MAE+DPPMask	**69.56**	**69.59**	**61.86**

Table 5. Multi-label classification accuracy on COCO.

Method	ACC	$F1_{macro}$	$F1_{micro}$	$F1_{weighted}$
MAE	80.69	44.81	45.45	46.46
MAE+DPPMask	**80.85**	**45.08**	**45.85**	**47.24**

Table 6. Multi-label classification accuracy on CLEVR.

ing strategy. To better understand the relationship between the masking ratio and the misalignment problem, we study the fine-tuned performance of MAE under different masking ratios. We perform each experiment three times and report their mean accuracy of ImageNet-100. For DPPs sampling, we run two values of τ, 0.90 and 0.85, we report the higher performance. As Fig. 4 shows, we find that the original 0.75 masking ratio is not the optimal setting for fine-tuned performance. Instead, lowering the masking ratio significantly improves the accuracy, this is further evidence of the impact of the misalignment problem on feature learning, as a higher masking ratio raises the probability of misalignment. With DPPs sampling equipped, our method has achieved higher performance in all masking ratio settings. In the masking ratio 0.7, the maximal performance boost reached 0.4%. When the masking ratio gets lower, the pretext task actually becomes more simple, which leads to a performance drop. Notably, we make a better result than the best in MAE (masking ratio at 0.6) with both higher and lower masking ratios (0.5, 0.7). This is meaningful, as the reconstruction problem becomes more simple while our method still performs better than MAE, which shows our sampling method makes the pre-train task more reasonable rather more simple.

4.3. Robustness

The misalignment problem makes the network align images with different semantics. In some severe cases, the network may are required to align the original image with the background. Thus, the misalignment problem can interfere with the network decision by letting the network more focusing the background of images. To verify this, we evaluate the quality of the learned feature on the background challenge [61]. We run fine-tuned models of each method on 4 different variations from the original image. Each variation replaces the original background with empty (O.F.), with another image in the same class (M.S.), with a random image in any class (M.R.), or with an image from the next class (M.N.), examples are shown in Fig. 5. Tab. 4 shows the performance of our method on the background challenge.

Both iBOT and MAE witnessed a steady performance gain in four different variations of original images. Notably, our method on MAE achieves 1.08% and 1.13% improvements on O.F. and M.N images respectively, which is higher than the original images (0.86%). Such results were also observed in the iBOT framework. Our method achieves 0.42% improvements on original images, while other variations both improved by a large margin. In particular, we achieve 1.58% improvements on M.S. which largely surpasses the original improvements. This shows our methods are more robust to background changes, as we do not impose the network to align the background to the original image as the random strategy does.

4.4. Multi-label classification

To further examine the influence of the misalignment problem, we also test our method on a multi-label classification task. An intuitive understanding is that: the network can not reflect the semantic changes, which are trained to reconstruct objects whether are been masked or not. Where in multi-label classification, every semantics is important, which makes them suitable to test whether the network is capable to extract whole information of images. We test our method on CLEVR [28] and MS-COCO [36] datasets, which are widely used in multi-label classification. The CLEVR dataset contains 24 binary labels, each indicating the presence of a particular color and shape (8×3) combination in the image [52]. For the MS-COCO dataset, we report the following fine-tuned performance on the validation set: mean average precision (mAP), average per-class F1 score ($F1_{class}$), and the average overall F1 score ($F1_{all}$) [50]. For the CLEVR dataset, we train MAE with and without DPPs for 200 epochs, we report the linear probing and fine-tuned performance of $F1_{micro}$, $F1_{macro}$ and $F1_{weighted}$, where 'micro' evaluate F1 across the entire dataset, 'macro' eval-

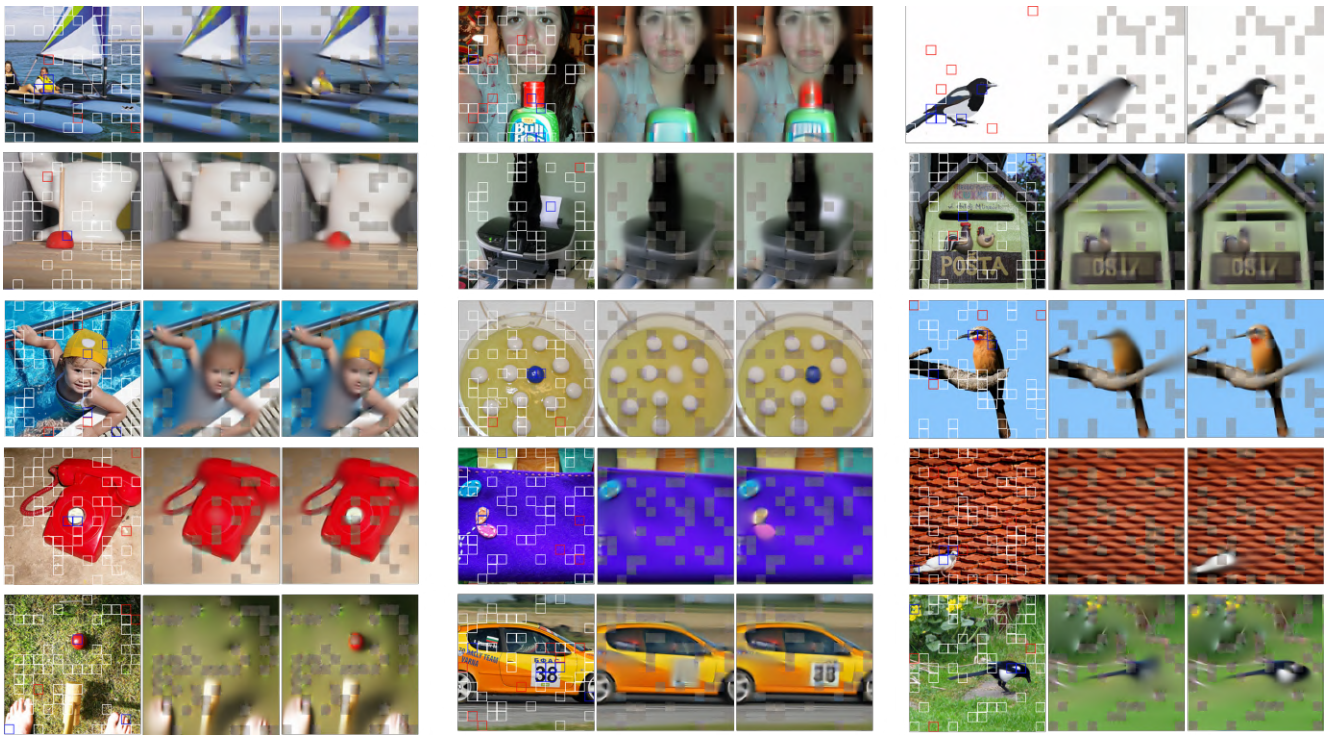

Figure 6. Comparison of DPPs sampling and random sampling, each triplet indicates the original image (right), reconstruction result with random sampling (middle), and DPPs sampling (left). The white boxes represent patches selected by both random and DPPs, while red boxes are for random sampling and blue boxes are for DPPs sampling. The threshold τ is set to 0.8.

uates an unweighted average of per-label F1 score, and 'weighted' scale the per-label F1 score by a number of examples when taking the average. We show our result in Tab. 5 and Tab. 6. Our method shows better performance on both COCO and CLEVR datasets. This shows our sampling method can select more informative patches for the network to reconstruct or align, which reduces the impact of the misalignment problem.

4.5. Qualitative analysis of DPPs sampling

To better understand the sampling behavior of DPPs, we compare the DPPs sampling result with random sampling. Fig. 8 shows the MAE reconstruction result from the ImageNet validation set, each triplet from left to right indicates the original image, reconstruction result with random sampling and DPPs sampling. For each image, we fix the random seed in order to find the difference between DPPs and random. We show coincide patches with white boxes, patches in random sampling while not in DPPs are shown in red boxes, and patches in DPPs while not in random are shown in blue boxes. Our experiment shows the reconstruction result of DPP sampling is better than random sampling, which proves that DPPs can represent more complete semantics than random. In particular, the sampling result shows two important properties of DPPs. First, DPPs can catch the appearance of each object more precisely, which is an important clue for image understanding. For example, the slot of the mailbox is crucial evidence to classify with cabin. Another important property is DPPs can retain more small foreground information, which is highly likely omitted in random sampling. Such properties show our method successfully alleviates the impact of misalignment problem, and achieve better performance in feature learning.

In our experiments, the MAE does not reconstruct the unobserved semantics, indicating that false positive samples are not perfectly aligned. A proper guess of such a phenomenon can be the diversity of ImageNet or the representative capabilities of networks. However, despite the network does not fall into over-fit, the incorrect gradient of misalignment will still interfere with the learning process. Our experiment also shows the potential of MIM with fewer misalignment problems.

5. Conclusion

In this paper, we show that uniformly random masking widely used in previous works unavoidably loses some key objects and changes original semantic information, resulting in a misalignment problem and hurting the representative learning. To this end, we propose a new masking strategy namely the DPPMask to reduce the semantic change of the image after masking. We show that DPPMask can make the MIM task more reasonable by purging the misalignment of training pairs. We hope our work can provide insights to help design a better MIM algorithm.

References

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

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3, 4, 12

[3] Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. A determinantal point process for column subset selection. *J. Mach. Learn. Res.*, 21:197–1, 2020. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 2

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 2

[7] 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 IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 2

[8] Laming Chen, Guoxin Zhang, and Eric Zhou. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in Neural Information Processing Systems*, 31, 2018. 5

[9] Mark Chen and Alec Radford. Rewon child, jeff wu, heewoo jun, prafulla dhariwal, david luan, and ilya sutskever. generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, volume 1, page 1, 2020. 3, 12

[10] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 3

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 2

[13] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. 12

[14] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 12

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[16] 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*, pages 1422–1430, 2015. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[18] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 2

[19] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[20] Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. DPPy: DPP Sampling with Python. *Journal of Machine Learning Research - Machine Learning Open Source Software (JMLR-MLOSS)*, 2019. Code at http://github.com/guilgautier/DPPy/ Documentation at http://dppy.readthedocs.io/. 4

[21] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[22] Jennifer Gillenwater, Alex Kulesza, and Ben Taskar. Near-optimal map inference for determinantal point processes. *Advances in Neural Information Processing Systems*, 25, 2012. 4

[23] Ross Goroshin, Joan Bruna, Jonathan Tompson, David Eigen, and Yann LeCun. Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 4086–4093, 2015. 2

[24] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017. 12

[25] Jie Gui, Zhengqi Liu, and Hao Luo. Good helper is around you: Attention-driven masked image modeling. *arXiv preprint arXiv:2211.15362*, 2022. 2, 6, 12

[26] 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*, pages 16000–16009, 2022. 1, 3, 5, 6, 12

[27] 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*, pages 9729–9738, 2020. 2

[28] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 7

[29] 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. 2, 12

[30] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[31] Chun-Wa Ko, Jon Lee, and Maurice Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995. 4

[32] Alex Kulesza, Ben Taskar, et al. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. 2, 3, 4

[33] Claire Launay, Agnès Desolneux, and Bruno Galerne. Determinantal point processes for image processing. *SIAM Journal on Imaging Sciences*, 14(1):304–348, 2021. 2

[34] Claire Launay and Arthur Leclaire. Determinantal patch processes for texture synthesis. In *XXVIIème Colloque GRETSI Traitement du Signal & des Images*, 2019. 3, 5

[35] Gang Li, Heliang Zheng, Daqing Liu, Bing Su, and Changwen Zheng. Semmae: Semantic-guided masking for learning masked autoencoders. *arXiv preprint arXiv:2206.10207*, 2022. 2, 6, 12

[36] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 7

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[38] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 12

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization (2017). *arXiv preprint arXiv:1711.05101*, 2019. 12

[40] Odile Macchi. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122, 1975. 3

[41] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European conference on computer vision*, pages 527–544. Springer, 2016. 2

[42] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14(1):265–294, 1978. 4

[43] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[44] Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2701–2710, 2017. 2

[45] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1, 3

[46] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 2

[47] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[48] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 1, 2

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[50] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 7

[51] 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. *International journal of computer vision*, 115(3):211–252, 2015. 6

[52] Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. Adversarial masking for self-supervised learning. In *International Conference on Machine Learning*, pages 20026–20040. PMLR, 2022. 2, 6, 7, 12

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 12

[54] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems*, 33:6827–6839, 2020. 2

[55] Nicolas Tremblay, Simon Barthelmé, and Pierre-Olivier Amblard. Determinantal point processes for coresets. *J. Mach. Learn. Res.*, 20:168–1, 2019. 5

[56] 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*, pages 1096–1103, 2008. 3

[57] 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. 1, 3

[58] Xiaolong Wang, Kaiming He, and Abhinav Gupta. Transitive invariance for self-supervised visual representation learning. In *Proceedings of the IEEE international conference on computer vision*, pages 1329–1338, 2017. 2

[59] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[60] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1910–1919, 2019. 2

[61] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 7

[62] 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*, pages 9653–9663, 2022. 1, 3, 6

[63] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 12

[64] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 12

[65] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *arXiv preprint arXiv:2210.08344*, 2022. 4, 6

[66] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 3

[67] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3, 4, 6, 12