Adversarial Masking for Self-Supervised Learning

Yuge Shi [1]   N. Siddharth [2]   Philip H.S. Torr [1]   Adam R. Kosiorek [3]

Abstract

We propose ADIOS, a masked image modeling (MIM) framework for self-supervised learning, which simultaneously learns a masking function and an image encoder using an adversarial objective. The image encoder is trained to minimise the distance between representations of the original and that of a masked image. The masking function, conversely, aims at maximising this distance. ADIOS consistently improves on state-of-the-art self-supervised learning (SSL) methods on a variety of tasks and datasets—including classification on ImageNet100 and STL10, transfer learning on CIFAR10/100, Flowers102 and iNaturalist, as well as robustness evaluated on the backgrounds challenge (Xiao et al., 2021)—while generating semantically meaningful masks. Unlike modern MIM models such as MAE, BEiT and iBOT, ADIOS does not rely on the image-patch tokenisation construction of Vision Transformers, and can be implemented with convolutional backbones. We further demonstrate that the masks learned by ADIOS are more effective in improving representation learning of SSL methods than masking schemes used in popular MIM models.

1. Introduction

The goal of Masked image modeling (MIM) is to learn image representations, in a self-supervised fashion, by occluding parts of the input images. MIM is inspired by significant advances in natural language modelling such as BERT (Devlin et al., 2019), where the model is trained to fill-in words randomly removed from a sentence (Fig. 1, top row). Recent work, including MAE (He et al., 2021) and BEiT (Bao et al., 2021), show that these gains are at least partially transferable to vision. The task of a MIM

Masked Language Model
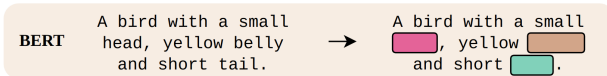
Masked Image Models

Figure 1. Self-supervised language, and vision, models learn representations by imputing data removed by masking. **BERT**: random word masks; **Context encoder**: random, fix-shaped mask; **BEiT**: random 'blockwise' masking; **MAE**: randomly mask out 75% of the image; **ADIOS**: multiple masks (N=3) generated by an adversarially trained masking model, post-processed with fully connected conditional random fields (Krähenbühl & Koltun, 2011).

model is therefore similar to BERT, e.g., given an image of a bird in Fig. 1, it needs to reason about what the bird might be sitting on or what colour the bird's belly is given the visible context (bottom row). However, while missing words describe *whole* semantic entities (e.g. "head"), the masks used for context encoder (Pathak et al. (2016), which pioneered MIM), BEiT and MAE typically have no such constraint (Fig. 1 bottom, left to right). Imputation under such schemes is conceptually simpler, as random masking only *partially* obscures meaningful visual entities, which allows easier inference of missing values by leveraging strong correlations at the local-pixel level[1].

To narrow the gap between pixel masking and word masking, we posit that one needs to occlude *whole entities in the image*. This encourages the model to perform imputation by complex semantic reasoning using the unmasked context (e.g. given a bird with a yellow body, it is likely to have a yellow head) rather than leveraging simple local correlations, which can benefit representation learning. Interestingly, He et al. (2021) are motivated by similar hypothesis and propose to occlude a large fraction (up to 75%) of the image, removing complete entities by a higher chance, which they find is essential for good performance. Here, we suggest that it is actually *what is masked*, not so much *how much is masked*, that is crucial for effective self-supervised

[1]University of Oxford [2]The University of Edinburgh & The Alan Turing Institute [3]DeepMind. Correspondence to: Yuge Shi <yshi@robots.ox.ac.uk>, Adam Kosiorek <adamrk@deepmind.com>.

---

[1]Akin to randomly masking letters in a sentence for NLP.

representation learning.

To this end, we investigate learning to mask with an adversarial objective, where an occlusion model is asked to make reasoning about missing parts of the scene more difficult. This novel representation-learning algorithm, called **Ad**versarial **I**nference-**O**cclusion **S**elf-supervision (ADIOS), can identify and mask out regions of correlated pixels within an image (Fig. 1, bottom right), which brings it closer to the word-masking regime in natural language. And as we shall see in Section 3, it consistently improves performance of state-of-the-art self-supervised learning (SSL) algorithms.

Some MIM methods employ a generative component for representation learning, by learning to reconstruct the masked image. However, it has been shown (Bao et al., 2021; Ramesh et al., 2021) that pixel-level reconstruction tasks waste modelling capacity on high-frequency details over low-frequency structure, leading to subpar performance. We hence frame ADIOS as an encoder-only framework that minimises the distance between the *representations* of the original image and the masked image. The occlusion model, which is trained adversarially to the encoder, tries to minimises this same distance. We further discuss in Section 2.1 that, compared to the generative setup, the encoder-only setup optimises a functionally superior objective for representation learning. Note that the encoder objective is compatible with many recent augmentation-based Siamese self-supervised learning (SSL; Chen et al. (2020); Chen & He (2021)) methods. We show that ADIOS consistently improves performance of these SSL objectives, showcasing the generality of our approach.

Our main contributions are as follows,

1. A novel adversarial Siamese-style MIM framework, that unlike other MIM methods is not limited to using ViT as backbone—advantageous given recent discoveries of modernised-convnet superiority over ViTs (Liu et al., 2022; Touvron et al., 2021);
2. Qualitative and quantitative analyses showing that masks generated by ADIOS are semantically meaningful;
3. Analysis of how different masking schemes affect representation-learning performance of SSL models. We find models trained with ADIOS and ground-truth object masks significantly outperform other masking schemes/no mask, demonstrating the efficacy of semantically meaningful masks for representation learning.

2. Methodology

Set up   ADIOS consists of two components, inference model $\mathcal{I}$ and occlusion model $\mathcal{M}$ (see Fig. 2). Given an RGB image $\boldsymbol{x}$, the occlusion model produces an image-sized mask $\boldsymbol{m} = \mathcal{M}(\boldsymbol{x})$ with values in $[0, 1]$. The inference model $\mathcal{I}$ takes original image $\boldsymbol{x}$ and an occluded image

$\boldsymbol{x^m} = \boldsymbol{x} \odot \boldsymbol{m}$ ($\odot$ is the Hadamard product) as inputs, generating representations for both, which we denote as $\boldsymbol{z}$ and $\boldsymbol{z}^m$. The two models are learnt by solving for

$$\mathcal{I}^\star, \mathcal{M}^\star = \arg\min_{\mathcal{I}} \max_{\mathcal{M}} \mathcal{L}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}). \quad (1)$$

We will now discuss different choices for $\mathcal{I}$ and $\mathcal{M}$.

2.1. Inference model $\mathcal{I}$

As discussed in Section 1, the inference model should minimise some distance between the original and masked images. Here, we discuss potential forms of this objective, arriving at our final framework using augmentation-based SSL methods.
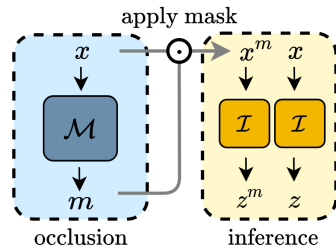

Figure 2. ADIOS Architecture.

Distance in pixel space   One option would be to inpaint the masked image with the inference model, and train $\mathcal{I}$ by minimising the distance between the inpainted image and the original image in pixel space. More specifically, we can define $\mathcal{I}$ as an auto-encoder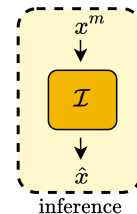 consisting of an encoder and decoder, which takes the masked image $\boldsymbol{x}^m$ as input and produces inpainted image $\hat{x}$ (see Fig. 3). The model can be trained using the following reconstruction loss

Figure 3. Inpainting.

$$\mathcal{L}_{\text{AE}}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) = \mathcal{D}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \mathcal{D}(\boldsymbol{x}, \mathcal{I}(\boldsymbol{x} \odot \mathcal{M}(\boldsymbol{x}))), \quad (2)$$

where $\mathcal{D}$ denotes some distance metric defined in pixel space. Minimising (2) encourages the auto-encoder to impute the missing part of the image as accurately as possible. $\mathcal{L}_{\text{AE}}$ can then be used in (1) to train the inference-occlusion model.

Distance in representation space   An interesting question for auto-encoding $\mathcal{I}$ is: where does the imputation happen? Multiple hypotheses exist: ① *Encoder only:* The encoder q completely recovers the missing information in the masked image, and in the ideal case, $q(\mathcal{M}(\boldsymbol{x})) = q(\boldsymbol{x})$; the decoder faithfully reconstructs these representations. ② *Decoder only:* The encoder faithfully extracts all information from the masked image, and the decoder reasons to inpaint missing pixel from the representations; ③ *Both:* The encoder and decoder both inpaint parts of the image.

Given these scenarios, ① is clearly best suited for representation learning, as it requires the encoder to reason about the missing parts based on observed context, beyond just extracting image features. With representation learning, rather than inpainting, being our end goal, the key challenge lies

in designing an objective targetting scenario ①, such that we learn the most expressive version of encoder q.

A key feature in ① is that when the encoder recovers all information of the original image, $q(\boldsymbol{x^m}) = q(\boldsymbol{x})$, the features extracted from the partially observed image $\boldsymbol{x^m}$ should in principle be the same as those extracted from the unmasked image $\boldsymbol{x}$. We thus propose an inference model $\mathcal{I}$ consisting of only the encoder, which extracts representation $\boldsymbol{z} = \mathcal{I}(\boldsymbol{x})$, $\boldsymbol{z} \in \mathbb{R}^d$. Our objective can thus be written as

$$\mathcal{L}_{\text{ENC}}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) = \mathcal{D}(\boldsymbol{z}, \boldsymbol{z^m}) = \mathcal{D}(\mathcal{I}(\boldsymbol{x}), \mathcal{I}(\boldsymbol{x} \odot \mathcal{M}(\boldsymbol{x}))) \quad (3)$$

where $\mathcal{D}$ is some distance metric defined in $\mathbb{R}^d$. Not only does $\mathcal{L}_{\text{ENC}}$ encourages the learning of more expressive encoder that can infer missing information, optimising this objective also does not involve generative component p, which is redundant for representation learning.

SSL framework (3) can be realised by simply optimising a Siamese network (Bromley et al., 1993). However, the objective can be trivially minimised when the representations for all
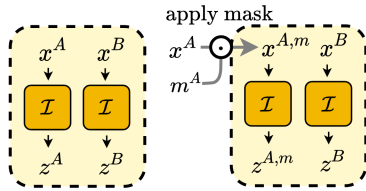

Figure 4. **Left**: SimCLR. **Right**: SimCLR + ADIOS.

inputs "collapse" to a constant. This phenomenon, known as latent collapse, has been addressed in many ways in augmentation-based SSL.

Let us take SimCLR (Chen et al., 2020) as an example (see Fig. 4, left); given a minibatch of M input images $\boldsymbol{x} = \{\boldsymbol{x}_i\}_{i=1}^M$, two sets of random augmentations A and B are applied to each image in $\boldsymbol{x}$, yielding $\boldsymbol{x}^A$ and $\boldsymbol{x}^B$. The same encoding function $\mathcal{I}$ is used to extract representations from both sets of augmented views, yielding $\boldsymbol{z}^A = \mathcal{I}(\boldsymbol{x}^A)$ and $\boldsymbol{z}^B = \mathcal{I}(\boldsymbol{x}^B)$. The objective of SimCLR is defined as

$$\mathcal{L}_{\text{SimCLR}}(\boldsymbol{x}; \mathcal{I}) = \log \frac{\exp(\mathcal{D}(\boldsymbol{z}_i^A, \boldsymbol{z}_i^B))}{\sum_{i \neq j} \exp(\mathcal{D}(\boldsymbol{z}_i^A, \boldsymbol{z}_j^B))}, \quad (4)$$

where $\mathcal{D}$ denotes the negative cosine similarity[2]. Intuitively, the objective minimises the distance between representations of the two augmented views of the same image (i.e. $\boldsymbol{z}_i^A, \boldsymbol{z}_i^B$), while repulsing the representations of different images (i.e. $\boldsymbol{z}_i^A, \boldsymbol{z}_j^B$). This effectively prevents the representations of different images from collapsing to the same constant, while optimising an objective similar to (3).

We can use the SimCLR objective for our model by masking one of the augmented images and then follow the exact same pipeline (see Fig. 4, right). More specifically, we replace $\boldsymbol{z}^A$ by $\boldsymbol{z}^{A,\boldsymbol{m}} = \mathcal{I}(\boldsymbol{x}^A \odot \boldsymbol{m}^A)$, where $\boldsymbol{m}^A = \mathcal{M}(\boldsymbol{x}^A)$ is a mask

---

[2] We omit the SimCLR temperature parameter for simplicity.

generated by the occlusion model given $\boldsymbol{x}^A$. Following (4), we can write the SimCLR-ADIOS objective as

$$\mathcal{L}_{\text{SimCLR}}^{\text{ADIOS}}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) = \log \frac{\exp(\mathcal{D}(\boldsymbol{z}_i^{A,\boldsymbol{m}}, \boldsymbol{z}_i^B))}{\sum_{i \neq j} \exp(\mathcal{D}(\boldsymbol{z}_i^{A,\boldsymbol{m}}, \boldsymbol{z}_j^B))}$$
$$= \log \frac{\exp(\mathcal{D}(\mathcal{I}(\boldsymbol{x}_i^A \odot \mathcal{M}(\boldsymbol{x}_i^A)), \mathcal{I}(\boldsymbol{x}_i^B)))}{\sum_{i \neq j} \exp(\mathcal{D}(\mathcal{I}(\boldsymbol{x}_i^A \odot \mathcal{M}(\boldsymbol{x}_i^A)), \mathcal{I}(\boldsymbol{x}_j^B)))} . \quad (5)$$

Again, we can use (5) in (1) to train the inference-occlusion model. Crucially, any SSL method that compares two augmented image views includes a term like (3), and can be plugged in to our framework. We conduct experiments using SimCLR, BYOL (Grill et al., 2020), and SimSiam (Chen & He, 2021) objectives and show significant improvement on downstream task performance with each method. Refer to Appendix A for the ADIOS objective used for BYOL and SimSiam, as well as more details on the SimCLR objective.

2.2. Occlusion model $\mathcal{M}$

For simplicity, we only consider the single-mask-generating case in the discussion above. In practice, since an image typically contains multiple components, we generate $N > 1$ masks to challenge the model to reason about relations between different components—empirical performance confirm benefits of doing so.

There are many parametric forms $\mathcal{M}$ could employ. For instance, one could consider generating multiple masks sequentially in an auto-regressive manner as seen in Engelcke et al. (2021). However, we find that the simplest setup suffices, where $\mathcal{M} : \mathbb{R}^{c \times w \times h} \mapsto \mathbb{R}^{N \times w \times h}$ consists of a learnable neural network and a pixelwise softmax layer σ applied across N masks to ensure that the sum of a given pixel across all the masks equals 1. We use U-Net as the backbone of our occlusion model—see Appendix B for more details. Note that we experimented with binarising the masks during training, but found that this did not yield improvements, and hence used real-valued masks directly.

2.3. Putting it together

Here we present ADIOS in its complete form. This includes the N-mask occlusion model, which generates masks $\{\boldsymbol{m}^{(n)}\}_{n=1}^N$ from the RGB image $\boldsymbol{x}$. The

Figure 5. ADIOS, $N > 1$.

inference model computes a loss $\mathcal{L}^{(n)}(\boldsymbol{x}; \mathcal{I}, \mathcal{M})$ for each $\boldsymbol{m}^{(n)}$ and the final loss is computed by averaging across N
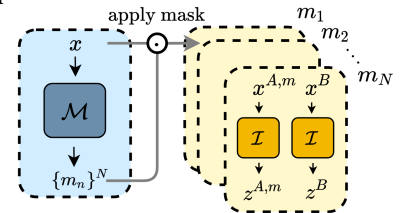
masks

$$\mathcal{I}^{\star}, \mathcal{M}^{\star} = \arg \min_{\mathcal{I}} \max_{\mathcal{M}} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}^{(n)}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) . \quad (6)$$

Sparsity penalty A trivial solution exists for this objective (6), where, for masks $\{\boldsymbol{m}^{(1)}, \ldots, \boldsymbol{m}^{(N)}\}$, some mask $\boldsymbol{m}^{(n)}$ occludes everything, with the other $\{N\}_{\backslash n}$ masks not occluding anything. To avoid such degenerate solutions, we introduce a sparsity penalty p_n in the form of $1/sin(\cdot)$ that discourages the occlusion model from generating all-one or all-zero masks; specifically,
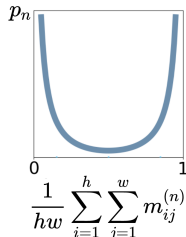

Figure 6. Penalty.

$$p_n = \sin\left(\frac{\pi}{hw} \sum_{i=1}^{h} \sum_{j=1}^{w} \boldsymbol{m}_{ij}^{(n)}\right)^{-1} . \quad (7)$$

Note that, p_n goes to infinity as $\boldsymbol{m}^{(n)}$ approaches all-one or all-zero (see Fig. 6). Minimising p_n with respect to $\mathcal{M}$ encourages the occlusion model to generate semantically meaningful mask, while avoiding degenerate solutions.

Final objective Let λ the scaling of the penalty term. Our complete objective reads as

$$\mathcal{I}^{\star}, \mathcal{M}^{\star} = \arg \min_{\mathcal{I}} \max_{\mathcal{M}} \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{L}^{(n)}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) - \lambda p_n\right) . \quad (8)$$

Lightweight ADIOS Despite its strong empirical performance, we note that the training objective in (8) requires N forward passes, which can be computationally expensive as we increase N for more complex data. We therefore develop a lightweight version of ADIOS, where we randomly sample one from the N generated masks to be applied to the input image. Doing so disassociates the computational cost of the model from the number of generated masks, and the only cost increase comes from applying the mask generation model once, which is inexpensive (10% the size of ResNet18). We name this single-forward pass version of our model ADIOS-s, and write the objective as

$$\mathcal{I}^{\star}, \mathcal{M}^{\star} = \arg \min_{\mathcal{I}} \max_{\mathcal{M}} \left(\mathcal{L}^{(k)}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) - \lambda \frac{1}{N} \sum_{n=1}^{N} p_n\right),$$
$$\text{where } k \sim \text{Uniform}\left(\{1, 2..., N\}\right). \quad (9)$$

3. Evaluation of Representations

Set up We evaluate ADIOS with three different SSL objectives: SimCLR (Chen et al., 2020), BYOL (Grill et al., 2020), and SimSiam (Chen & He, 2021). Each set of quantitative results is reported as an average over three random trials. We summarise our training setup in Appendix C.

Table 1. Top-1 classification accuracy (k-NN and Linear Probing) on Imagenet100-S, STL10. Improvements of ADIOS that are more than 3% are marked in bold.

Method	ImageNet100-S		STL10	
	k-NN	Linear	k-NN	Linear
Backbone: ViT-Tiny				
SimCLR	40.0 (± 0.28)	40.2 (± 0.47)	72.9 (± 0.27)	76.0 (± 0.33)
+ADIOS	42.0 (± 1.32)	43.1 (± 0.71)	73.4 (± 0.28)	**79.7** (± 0.88)
SimSiam	35.2 (± 1.12)	36.8 (± 1.82)	66.7 (± 0.10)	67.5 (± 0.02)
+ADIOS	**38.8** (± 2.73)	**40.1** (± 0.59)	67.9 (± 0.75)	68.8 (± 0.25)
BYOL	38.1 (± 0.61)	39.7 (± 0.50)	71.9 (± 0.12)	72.1 (± 0.32)
+ADIOS	**47.1** (± 0.35)	**49.2** (± 0.94)	74.5 (± 0.58)	**75.9** (± 0.63)
Backbone: ResNet-18				
SimCLR	54.1 (± 0.09)	55.1 (± 0.15)	83.7 (± 0.48)	85.1 (± 0.12)
+ADIOS	55.1 (± 0.43)	55.9 (± 0.21)	85.8 (± 0.10)	86.1 (± 0.36)
SimSiam	58.6 (± 0.31)	59.5 (± 0.31)	84.3 (± 0.81)	84.8 (± 0.72)
+ADIOS	61.0 (± 0.29)	60.4 (± 0.19)	84.6 (± 0.35)	86.4 (± 0.24)
BYOL	56.2 (± 0.79)	56.3 (± 0.10)	83.6 (± 0.09)	84.3 (± 0.13)
+ADIOS	**60.2** (± 0.82)	**61.4** (± 0.14)	84.8 (± 0.19)	85.6 (± 0.24)

3.1. Classification

We evaluate the performance of ADIOS on STL10, as well as a downsized version of ImageNet100 (Tian et al., 2020), from resolution 224x224 to 96x96. We refer to our version of the dataset as ImageNet100-S. We also evaluate the performance of ADIOS-s on the original ImageNet100 dataset. Both ImageNet100 and STL10 are derived from ImageNet-1k (Russakovsky et al., 2015): ImageNet100 contains data from 100 ImageNet classes, and STL10 is derived from 10 object classes of ImageNet, with 5,000 labelled images and 100,000 unlabelled images. Due to computational constraints we were unable to evaluate on ImageNet-1k; we leave this for future work.

For ADIOS, we provide results using ResNet18 (He et al., 2016) and ViT-Tiny (Dosovitskiy et al., 2021) backbones on three classification tasks: linear evaluation, k-NN and clustering. Through hyperparameter search, we use $N = 4$ masking slots for ImageNet100-S and $N = 6$ for STL10. For ADIOS-s we provide results using ResNet18 as backbone on linear evaluation.

Linear evaluation and k-NN We study the utility of learned representation by classifying the features using both linear and a k-nearest neighbour (k-NN) classifiers. Following the protocol in Zhou et al. (2021), we sweep over different numbers of nearest neighbours for k-NN and different learning rates for the linear classifier.

Results are presented in Tab. 1, where each +*ADIOS* entry represents the ADIOS framework applied to the SSL objective in the row above. For instance, the top coloured block shows results of SimCLR and SimCLR+ADIOS. Models using ADIOS consistently outperform their respective SSL baselines beyond the margin of error; in some cases achieving significant improvements of 3–9% (in bold).

Table 2. Top-1 classification accuracy of linear probing on ImageNet100. Improvements of more than 3% are marked in bold.

SimCLR	+ADIOS-s	SimSiam	+ADIOS-s	BYOL	+ADIOS-s
77.5 ($\pm$0.10)	76.1 ($\pm$0.50)	76.4 ($\pm$0.07)	77.2 ($\pm$0.09)	74.3 ($\pm$0.16)	**80.8 ($\pm$0.60)**

Notably, the ViT-Tiny models perform significantly worse than ResNet-18 models, which is unsurprising given that ViT-Tiny uses half the number of parameters of ResNet-18.

The best Top-1 accuracy on ImageNet100-S is 61.4%, achieved by BYOL+ADIOS using linear evaluation, surpassing its baseline BYOL by 5.1%; For STL10, the best performing model is SimSiam+ADIOS using linear evaluation with an accuracy of 86.4%, while SimSiam evaluates at 84.8%. Significantly, ADIOS improves all metrics on both backbones and both datasets. Interestingly, the degree of improvement varies by method, and can result in order change between respective SSL methods.

We also run experiments on the original ImageNet100 dataset with the single-forward pass version of our model, ADIOS-s. The results in Tab. 2 show that this much cheaper model also achieves impressive performance, especially when applied to BYOL with a performance boost of more than 6%. This result further demonstrates the efficiency of our approach, and the reduced computational cost allows for the potential of scaling to larger datasets.

Clustering Following Bao et al. (2021); Zhou et al. (2021), we also evaluate the trained models using standard clustering metrics, including adjusted random index (ARI) and Fowlkes-Mallows index (FMI), both of which computes the similarity between clusterings, as well as normalised mutual information (NMI). We assign pseudo-labels to the representation of each image using k-means, and evaluate the three metrics on the clusters formed by the pseudo-labels vs. the true labels. Results in Tab. 3 are consistent with our previous findings. ADIOS improves the performance of baseline SSL methods on all three metrics for both datasets.

Findings ADIOS significantly and consistently improves the quality of representation learned under a range of set-ups, across two datasets, two backbone architectures, three SSL methods and on five different metrics, highlighting the effectiveness and versatility of our approach.

3.2. Transfer learning

We study the downstream performance of models trained on ImageNet100-S, on four different datasets including CI-FAR10, CIFAR100 (Krizhevsky et al., 2009), Flowers102 (Nilsback & Zisserman, 2008), and iNaturalist (Horn et al., 2018). CIFAR10 and CIFAR100 resolutions are 32, while those of Flowers102 and iNaturalist are 96. We only use the

Table 3. Clustering performance on Imagenet100-S, STL10.

Method	Backbone	Metrics		
		FMI ↑	ARI ↑	NMI ↑
Dataset: ImageNet100-S				
SimCLR	ViT-Tiny	0.105 ($\pm$1e-3)	0.095 ($\pm$1e-3)	0.432 ($\pm$3e-3)
+ADIOS	ViT-Tiny	0.120 ($\pm$1e-3)	0.110 ($\pm$1e-3)	0.442 ($\pm$4e-3)
SimSiam	ViT-Tiny	0.077 ($\pm$9e-4)	0.067 ($\pm$2e-3)	0.389 ($\pm$3e-3)
+ADIOS	ViT-Tiny	0.098 ($\pm$1e-2)	0.087 ($\pm$9e-4)	0.425 ($\pm$3e-3)
BYOL	ViT-Tiny	0.098 ($\pm$8e-3)	0.088 ($\pm$8e-3)	0.418 ($\pm$4e-3)
+ADIOS	ViT-Tiny	0.132 ($\pm$3e-3)	0.123 ($\pm$1e-3)	0.458 ($\pm$4e-3)
SimCLR	ResNet18	0.151 ($\pm$3e-3)	0.135 ($\pm$4e-3)	0.515 ($\pm$6e-3)
+ADIOS	ResNet18	0.175 ($\pm$1e-3)	0.161 ($\pm$4e-3)	0.539 ($\pm$3e-3)
SimSiam	ResNet18	0.167 ($\pm$2e-3)	0.136 ($\pm$6e-3)	0.553 ($\pm$8e-3)
+ADIOS	ResNet18	0.179 ($\pm$1e-3)	0.161 ($\pm$1e-3)	0.553 ($\pm$1e-3)
BYOL	ResNet18	0.170 ($\pm$1e-3)	0.158 ($\pm$3e-3)	0.530 ($\pm$4e-3)
+ADIOS	ResNet18	0.179 ($\pm$6e-4)	0.156 ($\pm$2e-3)	0.561 ($\pm$2e-3)
Dataset: STL10				
SimCLR	ViT-Tiny	0.349 ($\pm$5e-3)	0.269 ($\pm$6e-3)	0.410 ($\pm$2e-3)
+ADIOS	ViT-Tiny	0.351 ($\pm$4e-3)	0.271 ($\pm$8e-3)	0.417 ($\pm$6e-3)
SimSiam	ViT-Tiny	0.296 ($\pm$3e-3)	0.177 ($\pm$1e-3)	0.341 ($\pm$4e-3)
+ADIOS	ViT-Tiny	0.320 ($\pm$3e-3)	0.235 ($\pm$5e-3)	0.349 ($\pm$0e-0)
BYOL	ViT-Tiny	0.349 ($\pm$5e-3)	0.269 ($\pm$5e-3)	0.410 ($\pm$5e-3)
+ADIOS	ViT-Tiny	0.355 ($\pm$4e-2)	0.276 ($\pm$3e-3)	0.422 ($\pm$4e-3)
SimCLR	ResNet18	0.338 ($\pm$2e-3)	0.166 ($\pm$9e-4)	0.512 ($\pm$5e-3)
+ADIOS	ResNet18	0.437 ($\pm$6e-3)	0.309 ($\pm$9e-3)	0.585 ($\pm$8e-3)
SimSiam	ResNet18	0.392 ($\pm$2e-3)	0.242 ($\pm$7e-3)	0.552 ($\pm$3e-3)
+ADIOS	ResNet18	0.412 ($\pm$8e-3)	0.249 ($\pm$7e-3)	0.558 ($\pm$2e-4)
BYOL	ResNet18	0.429 ($\pm$5e-3)	0.328 ($\pm$9e-3)	0.525 ($\pm$8e-3)
+ADIOS	ResNet18	0.508 ($\pm$6e-3)	0.422 ($\pm$1e-2)	0.588 ($\pm$9e-3)

ResNet-18 models here as they clearly outperform the ViT-Tiny models in our previous experiments. Detailed setup and hyperparameters are given in Appendix D.

Tab. 4 reports classification accuracy under two different transfer-learning setups, including *F.T*: fine-tuning the entire model and *Lin.*: freeze the encoder weights and re-train the linear classifier only. As a comparison, we also show the results of training from scratch on each dataset.

Results show that ADIOS improves transfer learning performance on all four datasets, under both linear evaluation and fine-tuning. Bigger improvements, of >3% (marked in bold) occur mostly under linear evaluation, indicating that compared to baseline SSL models, the ADIOS-pre-trained representations are much easier to linearly separate. Notably, all 6 models' fine-tuning performance exceeds training from scratch, demonstrating the benefits of pretraining.

One might also notice that different from the other datasets, there exists large discrepancies between the linear evaluation vs. fine-tuning performance on CIFAR10 and CIFAR100. This is because He et al. (2016) suggests a slightly different architecture for CIFAR with smaller kernel size in the first layer to suit its small image size (see details in Appendix F). We use CIFAR-ResNet for fine-tuning, however we have to use the original architecture for linear evaluation in order to use the pretrained weights, which leads to poor performance.

Table 4. Classification accuracy of transfer learning by re-training linear classifier only (*Lin.*) and fine-tuning (*F.T*). More than 3% improvements by ADIOS are marked in bold.

Method	CIFAR10		CIFAR100		Flowers102		iNaturalist	
	Lin.	F.T.	Lin.	F.T.	Lin.	F.T.	Lin.	F.T.
SimCLR	30.1	91.3	10.2	70.0	42.5	45.6	69.4	82.1
+ADIOS	**34.6**	93.4	11.0	71.8	**50.2**	**50.6**	**72.5**	84.3
SimSiam	35.3	92.4	13.2	65.0	38.7	55.0	72.3	85.0
+ADIOS	**39.3**	94.3	13.3	**71.0**	**44.9**	**59.0**	**75.9**	86.2
BYOL	29.9	88.0	13.3	52.3	49.1	58.6	72.7	85.1
+ADIOS	**39.2**	90.4	14.0	**62.0**	51.7	60.1	73.1	85.7
Scratch	-	85.5	-	49.8	-	30.6	-	73.8

3.3. Robustness

It is likely that the adversarial masks learned by ADIOS target informative image features, including spurious correlations if present in a dataset. It is therefore interesting to ask if ADIOS representations are robust to changes in such spurious correlations. To answer this, we evaluate models pretrained on ImageNet100-S on the backgrounds challenge (Xiao et al., 2021), where 7 different

Orig. **M.S.**

M.R. **M.N.**

Figure 7. Examples.

types of variation on a subset of ImageNet data are used to measure the impact of foreground and background on model decision making. Examples of such variations can be seen in Fig. 7, where the original figure's (*Orig.*) background is replaced by background from another image in the same class (*M.S.*), from a random image in any class (*M.R.*) or from an image in the next class (*N.R.*). We perform linear evaluation on the pretrained models on these variations, and report the classification accuracy in Tab. 5.

Our results show that all three SSL-ADIOS models outperform their respective baselines on *all* variations, demonstrating that ADIOS-learned representations are more robust to changes in both foreground and background.

It is useful to examine how ADIOS behaves in different testing conditions regardless of the SSL objective used. To this end, the bottom row of Tab. 5 contains the performance gain of ADIOS averaged over all three SSL models. We witness the biggest gains on the *M.R.* and *M.N.* conditions (bottom row, Fig. 7). That is, when any deterministic relation between labels and backgrounds is severed. The improvement is the lowest for the original images and *M.S.* condition (top row, Fig. 7), both of which preserve the relation between labels and background. This demonstrates that ADIOS depends less on background information that are spuriously-correlated with object labels when making predictions. This is not surprising—as we describe in Section 4.1—ADIOS'
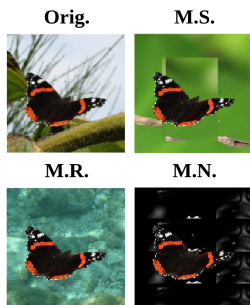
Table 5. Accuracy on different variations of the backgrounds challenge, evaluating model robustness. Example variations in Fig. 7.

Method	Variations							Orig.
	O.BB.	O.BT.	N.F.	O.F.	M.S.	M.R.	M.N.	
SimCLR	20.1	34.8	44.3	41.6	67.1	45.9	41.0	78.8
+ADIOS	20.7	36.7	45.5	43.5	68.0	47.9	43.7	79.1
SimSiam	29.5	39.1	43.8	52.1	69.9	43.9	40.8	78.4
+ADIOS	33.1	41.0	46.2	54.7	71.5	47.2	43.5	80.3
BYOL	25.9	38.4	46.0	51.6	71.3	45.6	42.7	79.8
+ADIOS	27.7	39.0	48.5	51.7	72.1	47.8	44.1	80.6
Avg. gain	+2.0	+1.4	+2.0	+1.5	+1.1	+2.5	+2.3	+1.0

generated masks tend to occlude backgrounds, encouraging the model to focus on the foreground objects.

4. Analysis on Learned Masks

Here, we look at the masks generated by ADIOS' occlusion model when trained on ImageNet100-S, STL10, and CLEVR (Johnson et al., 2017)—a dataset of rendered 3D objects such as cubes and spheres. We use $N = 4$ masking slots for CLEVR and include training details in Appendix E. The top row of each image block in Fig. 8 shows the original image. The bottom row displays the generated masks, with each colour representing one masking slot. See Section 4.1 for a detailed analysis. We also quantitatively analyse ADIOS' masks in Section 4.2.

4.1. Mask Generation

For realistic, single-object datasets such as STL10 and ImageNet100-S, ADIOS manages to mask out different compositions of the image. In the STL10 dataset, each image clearly shows a 'foreground' object and a 'background' (Fig. 8a). In this setting, ADIOS learns to mask specific object parts like the wings or the tail of a bird (4th column) and the mouth or the face of a horse (5th column) and the mouth or the face of a horse (5th column). In case of the ImageNet100-S dataset, however, it is often not obvious if the image features any particular entity. Hence, ADIOS tends to occlude complete entities like the animal and the ant in the in the 7th and 8th columns of Fig. 8c.

In CLEVR (simple rendered objects), ADIOS is usually able to put the background into a separate slot; the remaining slots split all present objects into 2-3 groups, with a tendency of applying a single mask to objects of the same colour (see Fig. 8b). While not a focus on our work, and in contrast to prior art (Greff et al., 2019; Engelcke et al., 2020), ADIOS does not produce perfect segmentations. It is however an interesting research direction, given that better segmentations could further improve representation learning performance as we show in Section 4.2.

Summary Qualitative results show that ADIOS can generate semantically-meaningful masks. Crucially, the generated masks focus on different levels of detail, depending on
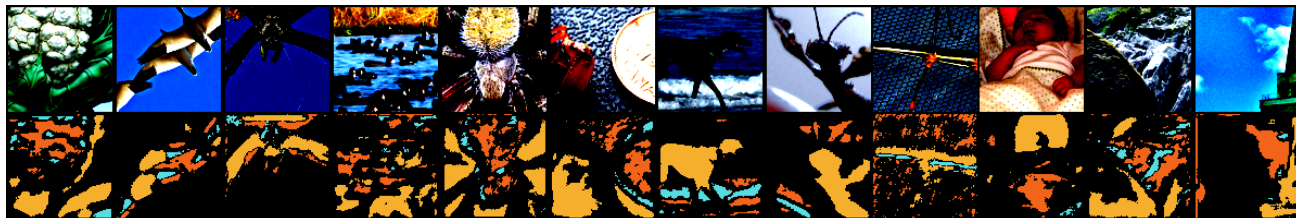
(a) STL10, $N = 6$.

(b) CLEVR, $N = 4$.

(c) ImageNet100-S, $N = 4$.

Figure 8. Masks generated by ADIOS during training on each dataset. N denotes the number of masks. *Top row:* original image; *Bottom row:* generated masks, each color represents one mask.

the dataset. This may explain some of the ADIOS' performance gains in the robustness experiments of Section 3.3, as semantic perturbations are baked into the training process.

4.2. Comparing Masking Schemes

The premise of our work is that for masked image modeling, *what is masked*, matters. In this section, we further investigate this by comparing the representation learning performance of SSL models trained under an ADIOS-like MIM framework, but with non-parametric masks including ground-truth semantic masks and random masks.

In Fig. 9 we outline the different masking schemes investigated in this section on the CLEVR dataset, including: a) ground-truth object segmentation masks (provided with the dataset), b) foreground-background masks (where the foreground is the union of all objects), c) ground-truth, box-shaped masks, d) shuffled ground-truth object segmentation masks (i.e. one image uses the ground-truth mask of another), e) random mask occluding 75% of the image, as in MAE (He et al., 2021), f) blockwise mask occluding 30% of the image, same as BEiT (Bao et al., 2021). Note that out of these masking schemes, a–c include semantic information, while d–f do not. In addition, we perform similar experiments on ImageNet100-S and STL10, however since we do not have information of the ground-truth object segmentation, we only compare the performance of ADIOS against using the random masking scheme in MAE and BEiT (i.e. e and f).

The representation learning performance of different masking schemes on ImageNet100-S and STL10 is evaluated by its top-1 classification accuracy under linear probing. To evaluate this on CLEVR, we set up a challenging multi-label classification task. Namely, we predict 24 binary labels, each indicating presence of a particular colour and shape (8×3) combination in the image. We report the F1 score (harmonic mean of the precision and recall) under different weighted average of subpopulation (defined by labels), where 'micro' evaluates F1 across the entire dataset, 'macro' evaluates an unweighted average of per-label F1 score, and 'weighted' scale the per-label F1 score by number of examples when taking average.

Tabs. 6 and 7 contains results averaged across these three SSL methods, each with three random trials (i.e. each entry in Tabs. 6 and 7 is averaged over nine runs). This is to marginalize out particular SSL methods, and therefore better understand effect of each of the masking schemes.
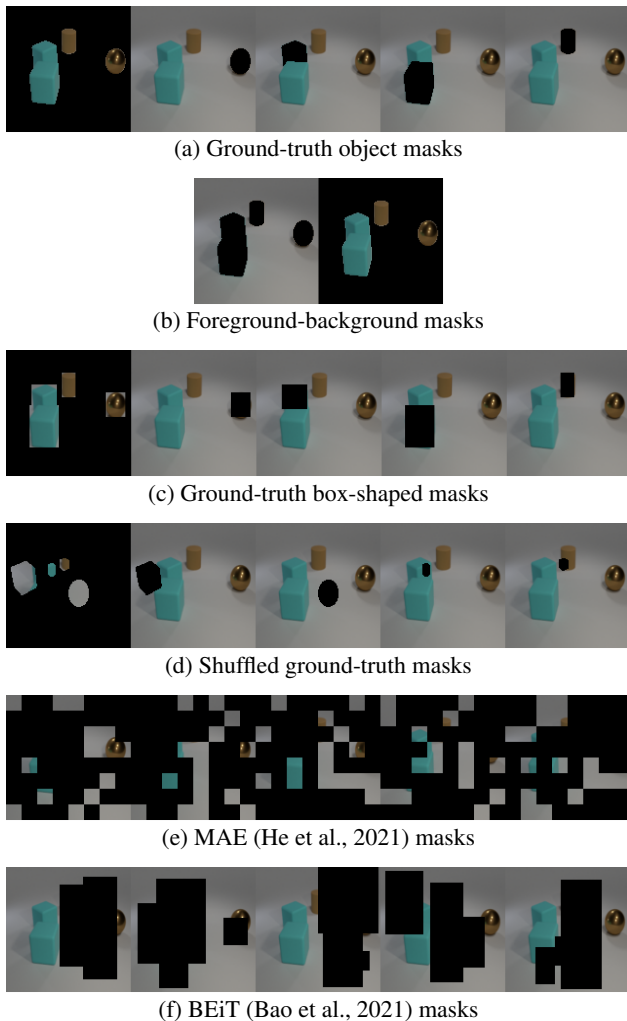
(a) Ground-truth object masks

(b) Foreground-background masks

(c) Ground-truth box-shaped masks

(d) Shuffled ground-truth masks

(e) MAE (He et al., 2021) masks

(f) BEiT (Bao et al., 2021) masks

Figure 9. Masking schemes used to compare against ADIOS.

The results clearly show the advantage of using semantic masks over non-semantic masks for representation learning. In Tab. 6, we show that ADIOS significantly outperform random masking schemes used in MAE and BEiT; additionally, the ground-truth object masks (*G.T.*) In Tab. 7 achieves the best performance on all three metrics, closely followed by ADIOS with comparable F1-macro, F1-weighted score, and slightly lower F1-micro score. We compare this to the randomly shuffled ground-truth mask (*Shuffle*), covering on average the same image fraction as the ground-truth, but where there is far less semantic consistency to the image content. The shuffled masks perform much worse on all three metrics, supporting our hypothesis from Section 1 that *what is masked* is more important that *how much is masked*.

The remaining masking schemes behave as expected, with the semantically-informed ones outperforming the non-semantic ones. Perhaps surprisingly, random masks, including the ones used in MAE and BEiT, can hurt representation learning under this ADIOS-like MIM framework: in both

Table 6. Top-1 classification accuracy on ImageNet100-S and STL10 under different masking schemes, averaged over three runs of SimCLR, SimSiam and BYOL respectively. Best results for each metric in bold.

Mask type	Condition	Dataset	
		ImageNet100-S	STL10
Random	e) MAE	43.7 (±0.43)	78.4 (±0.91)
	f) BEiT	46.4 (±0.67)	80.7 (±1.00)
Learned	ADIOS	**59.2** (±2.92)	**86.0** (±0.40)
None	-	57.0 (±2.27)	84.7 (±0.40)

Table 7. Multi-label classification on CLEVR under different masking schemes, averaged over three runs of SimCLR, SimSiam and BYOL respectively. Best results for each metric in bold.

Mask type	Condition	Metric		
		F1-macro ↑	F1-micro ↑	F1-weighted ↑
Semantic	a) G.T.	**0.373** ($\pm7e\text{-}3$)	**0.401** ($\pm2e\text{-}4$)	**0.460** ($\pm1e\text{-}2$)
	b) FG./BG.	0.346 ($\pm7e\text{-}3$)	0.365 ($\pm2e\text{-}4$)	0.402 ($\pm1e\text{-}3$)
	c) Box	0.347 ($\pm2e\text{-}4$)	0.391 ($\pm3e\text{-}5$)	**0.457** ($\pm5e\text{-}2$)
Random	d) Shuffle	0.332 ($\pm6e\text{-}3$)	0.360 ($\pm8e\text{-}4$)	0.418 ($\pm1e\text{-}3$)
	e) MAE	0.309 ($\pm8e\text{-}4$)	0.336 ($\pm3e\text{-}4$)	0.391 ($\pm9e\text{-}4$)
	f) BEiT	0.274 ($\pm1e\text{-}3$)	0.307 ($\pm2e\text{-}4$)	0.395 ($\pm7e\text{-}3$)
Learned	ADIOS	**0.377** ($\pm2e\text{-}3$)	0.385 ($\pm9e\text{-}4$)	**0.451** ($\pm1e\text{-}3$)
None	-	0.352 ($\pm9e\text{-}3$)	0.359 ($\pm2e\text{-}4$)	0.373 ($\pm2e\text{-}5$)

Tabs. 6 and 7, the performance of random masking schemes are much lower even compared to the baseline where no mask is applied. We do note that MAE and BEiT both contain many components beyond their masking schemes that are critical for their successful learning as reported in the respective works. These include image decoders, discrete-VAE tokeniser and the ViTencoder. Our evaluation focus exclusively on the masking schemes, and suggests that semantically meaningful masks lead to better representations, while random masks do not.

Summary   Our experiments here show two things. Firstly, that semantically meaningful masks can be used as an effective form of augmentation for SSL models, but the same cannot be said for random masks. And secondly, that the representations learned from using the masks generated by ADIOS are comparable in quality to those learned from using ground-truth object masks.

5. Related Work

Augmentation based SSL   Recent work has seen rapid development in SSL utilising image augmentation, with the core idea being that the representation of two augmented views of the same image should be similar. This involves a range of work that adopts a contrastive framework where the positive sample pairs (i.e. two views of the same image) are attracted and negative pairs (i.e. two views of different images) are repulsed (Chen et al., 2020; Gansbeke et al.,

2020; He et al., 2020; Chen et al., 2020; Caron et al., 2020), as well as non-contrastive approaches (Grill et al., 2020; Ermolov et al., 2021; Zbontar et al., 2021; Chen & He, 2021; Bardes et al., 2021) that are able to prevent latent collapse without negative pairs, which is considered to be computationally expensive.

Learning augmentations Several models have proposed to learn augmentation policies with supervision signals (Cubuk et al., 2019; Hataya et al., 2020) that is more favourable to the task at hand. Tamkin et al. (2021) applies these for SSL and proposes to learn perturbation to the input image with a *viewmaker* model that is trained adversarially to the main encoder network. Different from our work where masks are generated to occlude different components of the image, their learned perturbation is l_p-bounded and provides more "color-jitter" style augmentation to the input. Koyama et al. (2021) also proposes to learn mask-like augmentations by maximising a lower bound to the mutual information between image and representation while regularising the entropy of the augmentation. However their experiment is limited to an edited MNIST dataset, and they only consider learning augmentations for one SSL algorithm, SimCLR.

Masked image models More recently, models such as MAE (He et al., 2021), BEiT (Bao et al., 2021), iBOT (Zhou et al., 2021) have been motivated by masked language models like BERT, as we ourselves do, and achieved highly competitive results on SSL. All three methods make use of vision transformers and propose to "inpaint" images occluded by *random masks* in one way or another: MAE employs an autoencoder to inpaint images that are heavily masked, with the decoder discarded after pretraining. On the other hand, BEiT and iBOT both utilise tokenisers to first transform image patches into visual tokens, with BEiT using off-the-shelf pretrained tokenizer and iBOT training the tokeniser online. Similar to our work, rather than performing reconstruction in pixel space, they minimise the distance between the visual tokens of the complete image vs. the masked image. Recent work has also seen the application of random masks to modalities beyond vision including speech and text, achieving strong performance in all these domains (Baevski et al.). Our work is significantly different from all above since we employ *semantically meaningful masks* from the occluder model, jointly learned with the encoding model. Moreover, our model does not rely on additional components such as tokenisers/image decoder, and since the construction of the model does not require splitting the image into patches, is also not limited to using vision transformers as the backbone architecture.

6. Conclusion

We propose a novel MIM framework named ADIOS, which learns a masking function alongside an image encoder in an adversarial manner. We show, in extensive experiments, that our model consistently outperforms SSL baselines on representation learning tasks, while producing semantically-meaningful masks. We also provide detailed analysis on using different forms of occlusion as augmentation for SSL in general. We find that the best representation learning performance results from using semantically-meaningful masks, especially ground-truth ones, and that masks generated by the ADIOS' occlusion model are almost as good.

One caveat of our model is that the memory and computation cost increases linearly with the number of masking slots N, since the model requires N forward passes before each gradient update. This likely can be addressed by randomly sampling one mask for each forward pass, but is left to future work. Additionally, we want to investigate ADIOS performance on larger datasets such as ImageNet-1K or 22K and larger backbones like ViT-L, ViT-H. However, we believe that, as it is, ADIOS' strong performance on a variety of tasks under versatile conditions provides valuable insights on the design of masked image models. Future work on masked image modeling should consider not only objectives and model architectures, but also mask design— this work shows that semantic masks are significantly more helpful than random ones in aiding representation learning.

7. Acknowledgements

References

Baevski, A., Hsu, W.-N., Xu, Q., Babu, A., Gu, J., and Auli, M. Data2vec: A general framework for self-supervised learning in speech, vision and language. URL `https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language/`. Accessed: 2022-01-27.

Bao, H., Dong, L., and Wei, F. Beit: Bert pre-training of image transformers. *ArXiv preprint*, abs/2106.08254, 2021. URL `https://arxiv.org/abs/2106.08254`.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *ArXiv preprint*, abs/2105.04906, 2021. URL `https://arxiv.org/abs/2105.04906`.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., Lecun, Y., Moore, C., Säckinger, E., and Shah, R. Signature verification using a "siamese" time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4):669–688, 1993.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/70feb62b69f16e0238f741fab228fec2-Abstract.html`.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020. URL `http://proceedings.mlr.press/v119/chen20j.html`.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R. B., and He, K. Improved baselines with momentum contrastive learning. *ArXiv preprint*, abs/2003.04297, 2020. URL `https://arxiv.org/abs/2003.04297`.

Cubuk, E. D., Zoph, B., Mané, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 113–123. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00020. URL `http://openaccess.thecvf.com/content\_CVPR\_2019/html/Cubuk\_AutoAugment\_Learning\_Augmentation\_Strategies\_From\_Data\_CVPR\_2019\_paper.html`.

da Costa, V. G. T., Fini, E., Nabi, M., Sebe, N., and Ricci, E. Solo-learn: A library of self-supervised methods for visual representation learning, 2021. URL `https://github.com/vturrisi/solo-learn`.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423`.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Engelcke, M., Kosiorek, A. R., Jones, O. P., and Posner, I. GENESIS: generative scene inference and sampling with object-centric latent representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL `https://openreview.net/forum?id=BkxfaTVFwH`.

Engelcke, M., Jones, O. P., and Posner, I. Genesis-v2: Inferring unordered object representations without iterative refinement. *ArXiv preprint*, abs/2104.09958, 2021. URL `https://arxiv.org/abs/2104.09958`.

Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. Whitening for self-supervised representation learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3015–3024. PMLR, 2021. URL `http://proceedings.mlr.press/v139/ermolov21a.html`.

Gansbeke, W. V., Vandenhende, S., Georgoulis, S., Proesmans, M., and Gool, L. V. Scan: Learning to classify images without labels. In *ECCV (10)*, pp. 268–285, 2020.

Greff, K., Kaufman, R. L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., and Lerchner, A. Multi-object representation learning with iterative variational inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2424–2433. PMLR, 2019. URL `http://proceedings.mlr.press/v97/greff19a.html`.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. Á., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - A new approach to self-supervised learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html`.

Hataya, R., Zdenek, J., Yoshizoe, K., and Nakayama, H. Faster autoaugment: Learning augmentation strategies using backpropagation. In *16th European Conference on Computer Vision, ECCV 2020*, pp. 1–16, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL `https://doi.org/10.1109/CVPR.2016.90`.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. B. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL `https://doi.org/10.1109/CVPR42600.2020.00975`.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. *ArXiv*, abs/2111.06377, 2021.

Horn, G. V., Aodha, O. M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. J. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8769–8778. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00914. URL `http://openaccess.thecvf.com/content\_cvpr\_2018/html/Van\_Horn\_The\_INaturalist\_Species\_CVPR\_2018\_paper.html`.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C. L., and Girshick, R. B. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1988–1997. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.215. URL `https://doi.org/10.1109/CVPR.2017.215`.

Koyama, M., Minami, K., Miyato, T., and Gal, Y. Contrastive representation learning with trainable augmentation channel. *ArXiv preprint*, abs/2111.07679, 2021. URL `https://arxiv.org/abs/2111.07679`.

Krähenbühl, P. and Koltun, V. Efficient inference in fully connected crfs with gaussian edge potentials. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pp. 109–117, 2011. URL `https://proceedings.neurips.cc/paper/2011/hash/beda24c1e1b46055dff2c39c98fd6fc1-Abstract.html`.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Liu, Z., Mao, H., Chao-Yuan, W., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s. *arXiv preprint arXiv: 2201.03545*, 2022.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., and Efros, A. A. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2536–2544. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.278. URL `https://doi.org/10.1109/CVPR.2016.278`.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference*

on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 2021. URL `http://proceedings.mlr.press/v139/ramesh21a.html`.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Tamkin, A., Wu, M., and Goodman, N. D. Viewmaker networks: Learning views for unsupervised representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=enoVQWLsfyL`.

Tian, Y., Krishnan, D., and Isola, P. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Touvron, H., Cord, M., El-Nouby, A., Bojanowski, P., Joulin, A., Synnaeve, G., and Jégou, H. Augmenting convolutional networks with attention-based aggregation. *ArXiv preprint*, abs/2112.13692, 2021. URL `https://arxiv.org/abs/2112.13692`.

Xiao, K. Y., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=gl3D-xY7wLq`.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021. URL `http://proceedings.mlr.press/v139/zbontar21a.html`.

Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., and Kong, T. ibot: Image bert pre-training with online tokenizer. *ArXiv preprint*, abs/2111.07832, 2021. URL `https://arxiv.org/abs/2111.07832`.

A. ADIOS Objectives

In this section we detail the objective used to optimise ADIOS+SimCLR, ADIOS+SimSiam and ADIOS+BYOL.

A.1. SimCLR

We show a simplified version of SimCLR architecture in Fig. 4. In reality, the encoder $\mathcal{I}$ of SimCLR further factorises into two networks, including a base encoder $f(\cdot)$ which extracts the representation h used for downstream tasks, followed by a projection head $g(\cdot)$ which maps h to the final embedding that's used to compute the objective in (4).
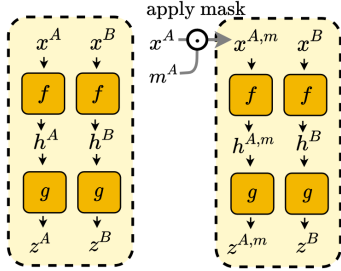
Figure 10. **Left**: SimCLR. **Right**: SimCLR + ADIOS.

We visualise this in Fig. 10. It is helpful to establish SimCLR's architecture as it lays the foundation for both SimSiam and BYOL.

A.2. SimSiam

Chen & He (2021) proposes SimSiam, an SSL method that can learn meaningful representations without negative examples. The forward pass of SimSiam also contains a base encoder and a projection head; however, different from SimCLR, for one of the augmentation streams the projection head is removed with a stop gradient operation applied to the base encoder (See Fig. 11). Authors find empirically that these two alterations are essential for preventing latent collapse in the absence of negative examples. The final objective of SimSiam is written as

$$\mathcal{L}_{\text{SimSiam}}(\boldsymbol{x}; \mathcal{I}) = \frac{1}{2} \left(\mathcal{D}(\boldsymbol{z}^A, \boldsymbol{h}^B) + \mathcal{D}(\boldsymbol{z}^B, \boldsymbol{h}^A) \right), \quad (10)$$

where $\mathcal{D}$ denotes the negative cosine similarity, $\mathcal{I} = g \circ f$, and $\boldsymbol{z} = g(f(\boldsymbol{x}))$ while $\boldsymbol{h} = f(\boldsymbol{x})$. Note that the loss is the average of two distances due to the asymmetrical model.
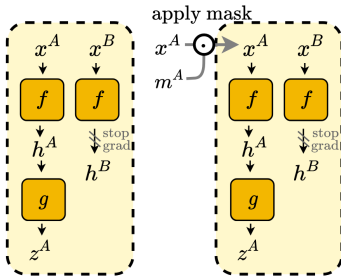
Figure 11. **Left**: SimSiam. **Right**: SimSiam + ADIOS.

Following the same intuition in Section 2, to adapt the objective for the ADIOS framework, we apply the masks learned by the occlusion model to one of the views.

We can therefore arrive at our final objective,

$$\mathcal{L}_{\text{SimSiam}}^{\text{ADIOS}}(\boldsymbol{x}; \mathcal{I}, \mathcal{M}) = \frac{1}{2} \left(\mathcal{D}(\boldsymbol{z}^{A,m}, \boldsymbol{h}^B) + \mathcal{D}(\boldsymbol{z}^{B,m}, \boldsymbol{h}^A) \right), \quad (11)$$

where $\boldsymbol{z}^{*,m} = g(f(\boldsymbol{x}^{*,m}))$.

A.3. BYOL

BYOL (Grill et al., 2020) is another SSL method that avoids the need of negative examples by performing an iterative online update. Similar to SimSiam, BYOL also adopts an assymetrical forward pass, however different from other approahces, the networks for two different augmentations do not share weights. See Fig. 12 for a visualisation.
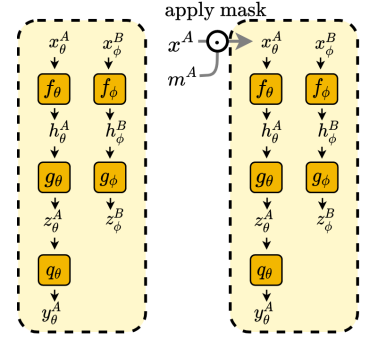
Figure 12. **Left**: BYOL. **Right**: BYOL + ADIOS.

For the sake of clarity, we denote the parametrisation of the two networks as θ and ϕ. The θ network is appended with an additional "predictor" q_θ, and is updated via gradient descent using the following objective, which is evaluated using the output of the θ network y_θ and output of the ϕ network z_ϕ

$$\mathcal{L}_{\text{BYOL}}(\boldsymbol{x}; \theta) = \frac{1}{2} \left(\mathcal{D}(\boldsymbol{y}_\theta^A, \boldsymbol{z}_\phi^B) + \mathcal{D}(\boldsymbol{y}_\theta^B, \boldsymbol{z}_\phi^A) \right), \quad (12)$$

where $\mathcal{D}$ denotes the mean squared error. Again, the objective is the average between the two terms due to the assymetrical architecture.

On the other hand, ϕ is optimised using the following update rule

$$\phi \leftarrow \tau \phi + (1 - \tau)\phi, \quad (13)$$

where $\tau \in [0, 1)$ controls the smoothness of the update.

To develop the ADIOS objective for BYOL, let us denote $\mathcal{I}$ as the composition of the two networks $\{\mathcal{I}_\theta, \mathcal{I}_\phi\}$. We can then write

$$\mathcal{L}_{\text{BYOL}}^{\text{ADIOS}}(\boldsymbol{x}; \mathcal{I}_\theta, \mathcal{M}) = \frac{1}{2} \left(\mathcal{D}(\boldsymbol{y}_\theta^{A,m}, \boldsymbol{z}_\phi^B) + \mathcal{D}(\boldsymbol{y}_\theta^{B,m}, \boldsymbol{z}_\phi^A) \right), \quad (14)$$

where $\boldsymbol{y}_\theta^{*,m} = q_\theta(g_\theta(f_\theta(x^{*,m})))$. Both $\mathcal{I}_\theta$ and $\mathcal{M}$ are optimised through the min-max objective in (6), whereas $\mathcal{I}_\phi$ is updated by (13).

Down

3x3 conv. 8 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 8 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 16 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 16 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 16 stride 1 pad 1 & GroupNorm & ReLU

MLP

F.C. 128 & ReLU
F.C. 128 & ReLU
F.C. 256 & ReLU

Up

3x3 conv. 16 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 16 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 8 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 8 stride 1 pad 1 & GroupNorm & ReLU
3x3 conv. 8 stride 1 pad 1 & GroupNorm & ReLU

Occlusion Head

1x1 conv. N stride 1 pad 1 & SoftMax

Table 8. U-Net Architecture.

Table 9. Hyperparameters for pretraining, used for all models.

Name	Value
Optimiser	SGD
Momentum	0.9
Scheduler	warmup cosine
Epochs	500
Batch size	128

Table 10. Learning rates for SimCLR, SimSiam and BYOL.

Architecture	Dataset	SimCLR	SimSiam	BYOL
ResNet18	ImageNet100-S	0.15	0.25	0.25
ResNet18	STL10	0.15	0.23	0.31
ViT-Tiny	ImageNet100-S	0.15	0.25	0.25
ViT-Tiny	STL10	0.15	0.11	0.23

Table 11. ADIOS hyperparameters for classification tasks.

	SimCLR+ADIOS	SimSiam+ADIOS	BYOL+ADIOS
Dataset: ImageNet100-S, backbone: ResNet18			
Enc. lr	0.13	0.85	0.24
Occ. lr	0.02	0.08	0.07
λ	0.57	0.29	0.40
N	4	4	4
Dataset: ImageNet100-S, backbone: ViT-Tiny			
Enc. lr	0.12	0.50	0.21
Occ. lr	0.03	0.07	0.33
λ	0.89	0.72	0.95
N	4	4	4
Dataset: STL10, backbone: ResNet18			
Enc. lr	0.21	0.52	0.49
Occ. lr	0.33	0.29	0.06
λ	0.29	0.79	0.72
N	6	6	6
Dataset: STL10, backbone: ViT-Tiny			
Enc. lr	0.14	0.56	0.29
Occ. lr	0.09	0.09	0.60
λ	0.50	0.12	0.18
N	6	6	6

B. Backbone of Occlusion Model

We use U-Net (Ronneberger et al., 2015) as the backbone of the occlusion model, which is commonly used for semantic segmentation. The model consists of a downsampling network, an MLP, and an upsampling network. We further apply an occlusion head layer with 1x1 kernel to map the output of U-Net to N masks. Refer to Tab. 8 for the architecture we used for our experiments.

C. Setups of Classificiation Tasks

We develop our model using solo-learn (da Costa et al., 2021), a library for state-of-the-art self-supervised learning methods. For backbones we use ResNet-18 (He et al., 2016) and ViT-Tiny (Dosovitskiy et al., 2021) with patch size 16x16. Hyperparameters including optimiser, momentum, scheduler, epochs and batch size are shared across all models, as seen in Tab. 9. We perform hyperparameter search on the learning rate of encoder for all models, and we include the optimal values used to generated the reported results in Tab. 10; for the ADIOS models we also run search for the learning rate of the occlusion model, the penalty scaling λ and number of masks N. Refer to Tab. 11 for the values used for these parameters.

D. Setups of Transfer Learning

We fine-tune all the models using SGD with a momentum of 0.9 with cosine learning rate decay. Following protocol in Dosovitskiy et al. (2021), we use batch size 512 and no weight decay. We also run a small grid search on the learning rate with values including $\{0.001, 0.003, 0.01, 0.03\}$.

E. Setups of CLEVR

CLEVR (Johnson et al., 2017) is a dataset of rendered 3D objects. The dataset contains detailed attributes for each object, including shape, color, position, rotation, texture as well as mask, and is commonly used in visual question answering and multi-object representation learning. Utilising the rich annotations of CLEVR, we construct a challenging multi-label classification task, which we use to evaluate the quality of representations learned under different masking scheme.

For hyperparameters including optimiser, momentum, scheduler, epochs and batch size, we follow the same setup in Tab. 9. We perform hyperparameter search on the learning rate of encoder and occlusion model, as well as the penalty scaling λ and number of masks N, which we list in Tab. 12.

Table 12. Hyperparameters used for CLEVR.

	SimCLR	+ADIOS	SimSiam	+ADIOS	BYOL	+ADIOS
Enc. lr	0.2	0.3	0.7	0.5	0.5	0.4
Occ. lr	-	0.1	-	0.1	-	0.3
λ	-	0.2	-	0.8	-	0.9
N	-	4	-	4	-	4

F. CIFAR ResNet

As we mentioned, the authors of ResNet (He et al., 2016) proposes a slightly different architecture for CIFAR due to their small image size. The difference between the CIFAR-ResNet and standard ResNet lies in the first convolutional block (before `layer1`), which we provide a side by side comparison of in Tab. 13 and Tab. 14.

Table 13. Standard ResNet.

7x7 conv. 64 stride 2 pad 3
BatchNorm
ReLU
3x3 MaxPool stride 2 pad 1

Table 14. CIFAR ResNet.

3x3 conv. 64 stride 1 pad 2
BatchNorm
ReLU
-

As we see, the kernel size of the convolutional layer is smaller and the MaxPool operation is removed to suit the small image size. We adopt this CIFAR-optimal ResNet for fine-tuning, but stick to the original architecture for linear evaluation to utilise the weights learned during pre-training, which resulted in the considerble underperformance on this metric for all 6 models.