Pseudo Labelling for Enhanced Masked Autoencoders

Srinivasa Rao Nandam[1]

Sara Atito[1, 2]

Zhenhua Feng[3]

Josef Kittler[2]

Muhammed Awais[1,2]

[1] Surrey Institute for People-Centred AI (PAI)
University of Surrey

[2] Centre for Vision, Speech and Signal Processing (CVSSP),
University of Surrey,
Surrey, UK

[3] School of Artificial Intelligence and Computer Science,
Jiangnan University,
Jiangsu, China

Abstract

Masked Image Modeling (MIM)-based models, such as SdAE, CAE, GreenMIM, and MixAE, have explored different strategies to enhance the performance of Masked Autoencoders (MAE) by modifying prediction, loss functions, or incorporating additional architectural components. In this paper, we propose an enhanced approach that boosts MAE performance by integrating pseudo labelling for both class and data tokens, alongside replacing the traditional pixel-level reconstruction with token-level reconstruction. This strategy uses cluster assignments as pseudo labels to promote instance-level discrimination within the network, while token reconstruction requires generation of discrete tokens encapturing local context. The targets for pseudo labelling and reconstruction needs to be generated by a teacher network. To disentangle the generation of target pseudo labels and the reconstruction of the token features, we decouple the teacher into two distinct models, where one serves as a labelling teacher and the other as a reconstruction teacher. This separation proves empirically superior to a single teacher, while having negligible impact on throughput and memory consumption. Incorporating pseudo-labelling as an auxiliary task has demonstrated notable improvements in ImageNet-1K and other downstream tasks, including classification, semantic segmentation, and detection.

1 Introduction

Masked Language modeling (MLM), introduced in Bert [13], has demonstarted that masking and predicting the masked tokens provides an effective pretext task for language. Maksed image modelling (MIM) introduced in SiT [1] has shown that masking large proportions of image regions/tokens and reconstructing the masked regions provides a strong pretext task

that enables self-supervised pretraining (SSP) in vision to outperform supervised pretraining (SP) and works well even in low data regime. BEiT [6], MAE [18], SimMIM [30] and several others [11, 19, 23] continued the idea of MIM, either at the pixel level or the token level, to develop state-of-the-art models that are effective across different downstream tasks.

Masked Auto Encoders [18] employ a sparse encoder where the masked regions are dropped and corresponding mask tokens are added only in decoder, typically consisting of 8 transformer blocks' to reconstruct the original image. The sparsity of the encoder makes MAE particularly appealing for pretraining of large and huge vision transformers. Consequently, various studies [10, 11, 15] have investigated various approaches to further enhance its performance. SdAE [11] replaces pixel level reconstruction of MAE with token level reconstruction. Evolved MAE [15] propose to adopt masking based on attention to further improve representations with MAE. CAE [10] adopts an auxiliary task of mask representation prediction which is similar to token level prediction in addition to reconstruction at pixel level utilised by MAE. To the best of our knowledge addition of auxiliary task which introduces a pseudo labelling task to MAE with dropping of tokens in encoders has not been yet explored. However, the addition of pseudo labelling or instance level discrimination with MIM methods without dropping of masked tokens for encoder has been explored by previous works [1, 3]. SiT [1] explored the incorporation of infoNCE [26] loss to the MIM pretext task, while MCSSL [3] investigated the addition of DINO loss on patches in conjunction with MIM through pixel level reconstruction and group masked model learning (GMML) [4]. Unlike MAE [18], these methods do not employ sparse encoders and utilise large amount of mask tokens (between 50% to 70%) for masking, which leads to inefficiencies in the pre-training stage. In contrast, MAE avoids this inefficiency by dropping the masked tokens, which is a significant portion of the input image ($\sim 75\%$).

In this work, we introduce a novel approach to Masked Image Modeling that combines the strengths of existing methodologies into a new framework. Our model features a sparse encoder similar to MAE, yet performs reconstruction in the token embedding space, akin to strategies used in [10, 11]. This design choice enables more contextually rich information capture. We further enhance our model's discriminative capability by integrating an auxiliary pseudo-labeling pretext task. Pseudo labels are generated through clustering by a dedicated pseudo-labeling teacher, enriching the model with nuanced, discriminative information.

Distinctively, our approach diverges from other multi-pretext task models such as [1, 3], and traditional instance discrimination methods [2, 5, 7, 31], by employing two specialised teachers for each pretext task. The employment of dual teachers effectively disentangles the outputs from each task, ensuring cleaner, more distinct learning signals. Additionally, we utilize different augmented views for different pretext tasks, sourced from two distinct augmentation strategies: one simple [18] and one complex [7]. This leads to misalignment between the inputs of the pseudo labelling teacher and the student. To resolve this, we address the issue by selecting the most similar patch in the feature space between the teacher and the student. The cluster from the most similar teacher patch is then considered as the target for clustering loss for each masked student cluster prediction. Inspired from SdAE [11], which introduces multifold masking to maintain similar mutual information between teacher and student similar, our model extends this strategy to the pseudo-labeling task. This adaptation demonstrated a noticeable improvement in the downstream tasks. Figure 1 provides a comprehensive overview of our innovative Pseudo MAE architecture, which effectively and efficiently capturing both fine-grained details and descriminative features.

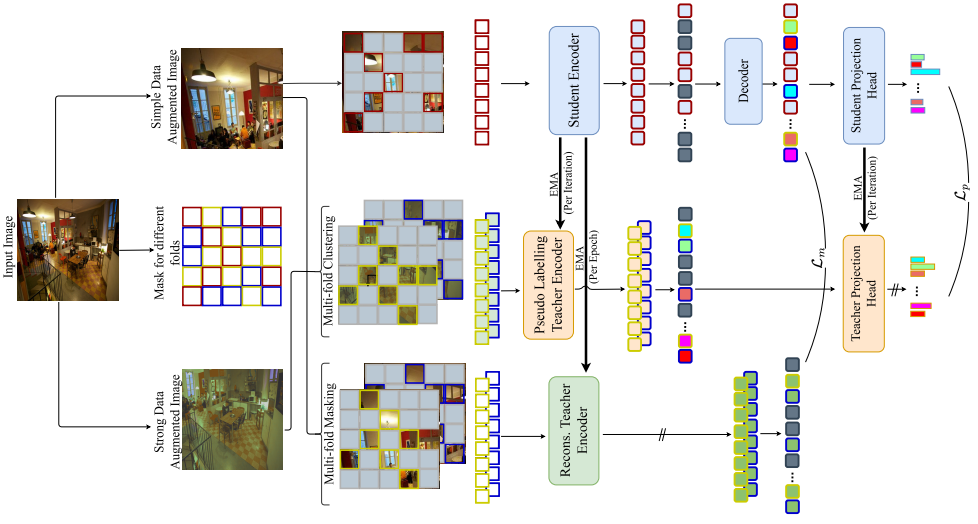

Figure 1: The proposed architecture consists of a student encoder, a decoder, and two disentangled teacher encoders updated using EMA. The student encoder processes unmasked regions while the masked tokens are reintroduced at the student decoder. Input image tokens are shuffled via random permutation, with each quarter allocated differently. Multifold pseudo labelling loss is applied to unaligned patch tokens of student and teacher based on feature similarity. The class token pseudo labelling loss is omitted for simplicity.

2 Related Work

Recent methods for SSL are based on either contrastive/clustering or MIM methods. Contrastive/Clustering methods focus on global semantic representation and can be broadly classified as instance discrimination methods. BERT [13] introduced the concept of masking random tokens in a transformer input and predicting those masked tokens for language. For computer vision the first MIM method GMML, which outperformed supervised pretraining, has been introduced in SiT [1]. It considers random groups of tokens from an image for masking and predicts the representation in pixel space. Following GMML, SimMIM [30] also uses MIM in pixel space with a light weight decoder. MAE [18] explored the concept of sparse encoder where the masked tokens are dropped and reconstruction at pixel level is done with the help of a deep transformer based decoder by reintroducing dropped tokens. BeiT [5] explored MIM through reconstruction at token level using codebook. The codebook for the target token representation is generated by a pretrained dVAE which has to be trained separately. However, BeiT could not outperform MIM methods which used simple pixel level reconstruction. For large and huge ViTs MAE is compute efficient compared to other methods like [1, 5] because of sparse processing of only visible tokens in encoder.

Several methods [10, 11, 15, 19, 20, 23] have proposed enhancements to MAE by adding different components or applying it to various hierarchical architectures. UM-MAE [20] proposes a two stage masking strategy for making MAE pretraining viable for hierarchical architectures like PVT[27], Swin[24]. GreenMIM[19] applies MAE on hierarchical models with the help of a new window attention mechanism and dynamic programming. MixMAE[23] proposes an simple strategy for applying MAE to hierarchical models by considering mixing of multiple images as masking. CAE[10] separates the pretraining tasks from learning

representations with the help of an additional alignment module. [15] proposes a masking strategy based on the attention maps of the encoder to capture better semantics. This new masking strategy learns objects parts effectively further enhancing MAE.

Addition of auxiliary contrastive task to enhance MIM has been studied even in the first successful ViT based MIM model SiT [1]. SiT [1] proposed GMML which reconstructs the image in original pixel space and considers addition of constrastive loss [9] as an auxiliary task to enhance the representations. They aim to capture finegrained information with the GMML and global information with contranstive auxiliary task. MCSSL [3] propose a framework which captures information at different levels. They use GMML to capture finegrained information and extend DINO [7] based clustering loss to patches to capture local semantics. iBoT [31] proposes to extend DINO to patches but does not reconstruct masked images at pixel level.

Different from these methods Pseudo MAE is different from other works by introducing auxiliary pseudo labelling pretext task to a sparse encoder based MAE where the main pretext task is MIM. We also change reconstruction target from pixels to tokens. We also extend the concept of multifold masking and propose multifold pseduo labelling. We also differ from instance discrimination based methods by different type of augmentations for both teacher and student. Differently from MCSSL and iBoT our method does not need multiple local and global crops and hence the training speed is significantly higher while maintaining performance advantage.

3 Methodology

Our method is an auto-encoder model where the sparse student encoder is a ViT [14] and the decoder consists of a few transformer blocks. The student decoder generates token embeddings at the output of the decoder which are used directly for masked reconstruction loss $\mathcal{L}_m$ in feature space. The same token embeddings at the output of the decoder are then passed through projection head to generate pseudo labels. The targets for reconstruction and pseudo labelling are generated by two disentangled teacher encoders. The pseudo labelling teacher generates pseudo labels from the outputs of a projection head, and the token teacher acts as a code-book which generates target tokens. The input to the pseudo labelling teacher and reconstruction teacher are generated from different augmentations of the original input image. The student and teacher share similar input view which is generated from simple data augmentation used by MAE [13]. The input view for pseudo labelling teacher is generated from a different crop of the original input based on harsher data augmentations used in DINO [7]. We use a multifold masking strategy introduced by SdAE [11] to reconstruction teacher, in addition we introduce a multifold class and patch level pseudo labelling which compares class pseudo labels across multiple folds. The class level pseudo labelling enables the network to learn instance level discrimination which is absent in other MIM based methods like MAE[13], SimMIM [30]. Our patch level pseudo labelling learns local patch level discrimination which is different from our token level reconstruction which captures global shape information similar to SdAE [11]. We discuss the framework and the different components in detail in the following sections.

3.1 Masked Image Modelling

Masked Image Modelling (MIM) has emerged as a significant self-supervised learning approach in recent research [1, 4, 11, 18, 30], showcasing its effectiveness and broad utility in the domain of image processing. At its core, MIM revolves around the reconstruction of an original image from an input image that has been partially masked. In this section, we provide an in-depth discussion of our novel MIM approach, utilising vision transformers.

Consider an input image $\mathbf{X}_{in} \in \mathcal{R}^{H\times W\times C}$, where H, W, and C denote height, width, and the number of channels, respectively. We generated two different $\mathbf{X}_{simp}, \mathbf{X}_{comp}$ views based on data augmentations used in MAE [18] and DINO [7] respectively. Both reconstruction teacher and student use the $\mathbf{X}_{simp}$ as their input, whereas pseudo labelling teacher uses $\mathbf{X}_{comp}$. The image is then converted into a sequence of flattened patches, resulting in a two-dimensional matrix $\mathbf{X}_{simp} \in \mathcal{R}^{N\times(P^2.C)}$, where N is the number of patches and $P\times P$ is the resolution of each patch. Subsequently, we generate a random mask $\mathbf{M} = (m_1,\ldots,m_N)$ for the N tokens, with each $m_i \in \{0,1\}$, determining the tokens to be retained or discarded.

For the student, we generate the masked input $\hat{\mathbf{X}} \in \mathcal{R}^{(N'+1)\times(P^2.C)}$, where N' is the number of the tokens corresponding to $m_i = 0$ excluding the class token. The output of the student after passing through the encoder and decoder which adds random mask tokens for masked regions is $\tilde{\mathbf{X}} \in \mathbf{R}^{(N+1)\times D}$, where D is the output embedding dimension. The objective function for reconstruction is simply a cosine similarity loss between targets and decoder output embeddings.

$$\mathcal{L}_m = \frac{1}{(N-N')}\sum_{i=1}^{N} cos\_sim(m_i \times (\overline{t_{rec}(\mathbf{X}_i^u)}, \tilde{\mathbf{X}}_i))\forall i \quad \text{where} \quad m_i = 1 \tag{1}$$

$$\overline{t_{rec}(\mathbf{X}_i)} = \frac{t_{rec}(\mathbf{X}_i) - mean(t_{rec}(\mathbf{X}_i))}{\sqrt{var(t_{rec}(\mathbf{X}_i)) + \varepsilon}} \tag{2}$$

Equation 1 gives the equation for reconstruction loss where t_{rec} is the reconstruction teacher encoder generating target tokens excluding class token. Here $\mathbf{X}_u$ is the input to the reconstruction teacher. This is different from MAE which uses original input pixels as the targets. Here $\tilde{\mathbf{X}}_i = (\tilde{\mathbf{X}}_1,\ldots,\tilde{\mathbf{X}}_N)$ is the output corresponding to patches excluding the class token output. We use normalised teacher inputs as the targets which are given by the Equation(2).

In addition we use a multifold masking strategy [11] for generating targets with teacher encoder. Let $\mathbf{X}_u$ represent the input where $\mathbf{X}_u = \{\mathbf{X}_{simp}(i) \mid \mathbf{m}_i = 0\}$. The input $\mathbf{X}_u \in \mathbf{R}^{(N-N')\times(P^2.C)}$ is split in K folds $(\mathbf{X}_u^1,\ldots,\mathbf{X}_u^k)$ to which class tokens are prepended to generate multifold split outputs $((t_{rec}(\mathbf{X}_u^1),\ldots,t_{rec}(\mathbf{X}_u^k))$. The output patches excluding that of class token of these multifold outputs is utilised in Equation(1).

3.2 Disentangled Teachers

Several works like [1, 3, 31] explore combining their primary pretext task clustering/contrastive learning with auxiliary pretext task of MIM. Apart from their encoders not being sparse due to the usage of mask tokens/zeros, they utilise a single teacher for both clustering/contrastive learning and their auxiliary MIM task which is updated with exponential moving average (EMA). We propose to use two teachers, namely pseudo labelling teacher t_{cl} and reconstruction teacher t_{rec}, that disentangle the task of generating target cluster assignments from

generating target token representations. The teacher for token representation benefits from having stable teacher where the teacher is updated by EMA less frequently, i.e. per epoch. The teacher for target cluster assignments performs better when updated frequently, i.e. per iteration. Utilising two different teachers specialised in different pretext tasks enables us to also apply two different EMA update schedules with different frequencies. Specifically, the EMA for our pseudo labelling teacher is done for each iteration with a starting momentum of 0.996 where as the EMA of the reconstruction teacher is done for each epoch with a starting momentum of 0.96.

3.3 Multifold Pseudo Labelling

Multifold masking produces noticeable improvements for downstream tasks as shown in [11]. We aim to extend multifold masking strategy for pseudo labelling as well. We generate pseudo labelling unmasked input $\mathbf{X}_{cu} = \{\mathbf{X}_{comp}(i) \mid \mathbf{m}_i = 0\}$ from complex data augmented view X_{comp} with mask m. Given the pseudo labelling unmasked input $\mathbf{X}_{cu} \in \mathbf{R}^{(N-N')\times(P^2.C)}$, we split in K folds $(\mathbf{X}^1_{cu}, \ldots, \mathbf{X}^k_{cu})$ and prepend class token to each fold. The resulting folds are passed through pseudo labelling teacher t_{cl} to generate multifold pseudo labelling split outputs $((\bar{\mathbf{X}}^1, \ldots, \bar{\mathbf{X}}^k)$ which include class and patch teacher encoder outputs for each fold. These encoder outputs are then passed through the projection head with a few linear layers and has a final output projection dimension of 4096 for both patch and class token layers. The pseudo labels are generated as cluster assignments from projection head outputs. We apply sinkhorn normalization to avoid collpase and generate target cluster predictions. Let $(\mathbf{P}^1, \ldots, \mathbf{P}^k)$ be the target patch pseudo labels and $(\mathbf{C}^1, \ldots, \mathbf{C}^k)$ be the target class pseudo labels after projection head for different folds.

We generate student decoder predictions $\tilde{\mathbf{X}}^m = \{\tilde{\mathbf{X}}_i \mid \mathbf{m}_i = 0\}$ corresponding to masked regions. $\tilde{\mathbf{X}}^m$ is passed through the student projection head to generate pseudo labels $\tilde{\mathbf{P}}, \tilde{\mathbf{C}}$ corresponding to patch and class tokens respectively. We generate inputs for pseudo labelling and student based on two different augmentations which causes misalignment between these views. This makes applying clustering loss non trivial unlike other clustering methods [7, 31] where similar views are passed through both teacher and student. To apply clustering loss on unaligned patches, we find the closest target pseudo label for each student pseudo label predictions. We perform the closest match by finding nearest teacher patch for each student patch in the feature space. Let $\hat{\mathbf{P}}_n = \mathbf{P}^k_i(\ where\ \mathbf{P}^k_i = min\{dist(\tilde{\mathbf{X}}_n, \bar{\mathbf{X}}^k_i)\})$ be the closest target pseudo label for n^{th} student label prediction. For the final class token target $\hat{\mathbf{C}} = Avg(\mathbf{C}^k)$ pseudo label, we take the average for all folds.

$$\mathcal{L}_p = \frac{1}{N-N'} \sum_{n=1}^{N-N'} \mathrm{H}(\hat{\mathbf{P}}_n, \tilde{\mathbf{P}}_n) \tag{3}$$

$$\mathcal{L}_c = \mathrm{H}(\hat{\mathbf{C}}, \tilde{\mathbf{C}}) \tag{4}$$

To calculate the patch level pseudo labelling loss, we apply cross entropy loss H between different teacher patch folds $\mathbf{p}^i$ and student patch folds $\tilde{\mathbf{p}}^i$ which is given by Equation 3. For class level pseudo labelling loss, we calculate loss between the single student class pseudo label $\tilde{\mathbf{c}}$ with teacher class pseudo label of each fold $\mathbf{c}^i$. The Equation 4 provides formulation for class teacher class pseudo labelling loss.

$$\mathcal{L} = \lambda_m * \mathcal{L}_m + \lambda_c * \mathcal{L}_c + \lambda_p * \mathcal{L}_p \tag{5}$$

The final loss is given by Equation 5, is the combination of reconstruction loss $\mathcal{L}_m$, class level pseudo labelling loss $\mathcal{L}_c$, and patch level pseudo labelling loss $\mathcal{L}_p$ with scaling factors $\lambda_m, \lambda_c, \lambda_p$ respectively set to 1 by default.

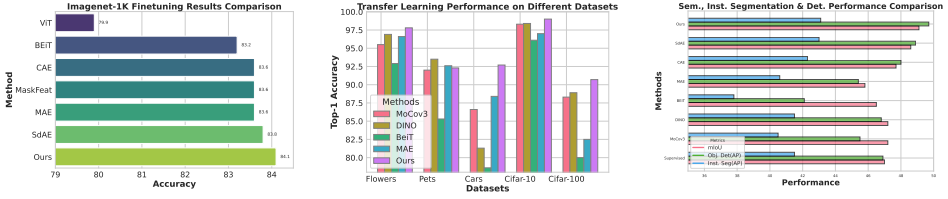

Table 1: Plots for (a)ImageNet classification, (b)Transfer Learning, (c)Det. & Seg.

4 Experiments

We base our experiment setup on pretraining first and finetuning next to evaluate our approach across several downstream tasks and datasets. We pretrain our model based on Imagenet-1K[12] following previous works[7, 18, 31]. We evaluate our model on multiple downstream tasks like multi class classification, semantic segmentation, objection detection, instance segmentation. The pretraining for Imagenet-1K is done for 800 epochs with AdamW[25] as the optimiser. The default image size is set to 224×224 and the patch size for ViT[14] is set to 16. The masking ratio is set to 75% following MAE, the learning rate is set to $2.666e-4$ and warmup epochs to 60. The staring, ending momentum for momentum schedule of reconstruction teacher is set to $0.96, 0.99$ respectively where the momentum update is done for each epoch. The staring, ending momentum for momentum schedule of pseudo labelling teacher is set to $0.996, 1$ respectively where the momentum update is done for each iteration. More ablations and visualisation to be provided in the supplementary section. Performance plots provided in Table 1

4.1 Results on multi class classification

We evaluate our model on Imagenet-1K[12] by finetuning for 100 epochs on ViT-B in Table 2. We set a base learning rate of $5e-4$ warmup epochs to 5. We utilise an effective batch

Method	Backbone	Supervision	Pretrain Epochs	Finetuning
ViT[14]	ViT-B	RGB	-	79.9
BEiT[3]	ViT-B	DALL-E	800	83.2
CAE[6]	ViT-B	DALL-E	800	83.6
MaskFeat[32]	ViT-B	HOG	300	83.6
iBOT[36]	ViT-B	Momentum	1600	84.0
MCSSL[8]	ViT-B	Momentum	800	84.0
MAE[18]	ViT-B	Momentum	800	83.6
SdAE*[4]	ViT-B	Momentum	300	83.8
Ours	ViT-B	Momentum	300	**84.1**

Table 2: Imagenet-1K finetuning results on various methods compared with ours. * represents our replicated model of SdAE with official pretraining and finetuning code.

Method	Backbone	Flowers	Pets	Cars	Cifar-10	Cifar-100
MoCov3	ViT-B	95.5	92.0	86.6	98.3	88.3
DINO	ViT-B	96.9	93.5	81.3	98.4	88.9
BeiT	ViT-B	92.9	85.3	78.6	96.1	80.0
MAE	ViT-B	96.6	**92.6**	88.4	97.0	82.5
Ours	ViT-B	**97.8**	92.3	**92.7**	**99.0**	**90.7**

Table 3: Transfer learning by finetuning pretrained models with the ViT-B/16 backbone on diverse datasets. We report top-1 accuracy and the results for different methods are directly from [3].

Method	Backbone	Pretrain Epochs	FLOPs (G)	Params (M)	mIoU
Supervised	ViT-B	300	606	164	47.0
MoCov3	ViT-B	300	606	164	47.2
DINO	ViT-B	400	606	164	47.2
BEiT	ViT-B	300	606	164	45.5
BEiT	ViT-B	800	606	164	46.5
MAE	ViT-B	300	606	164	45.8
MAE	ViT-B	1600	606	164	48.1
CAE	ViT-B	300	606	164	47.7
SdAE	ViT-B	300	606	164	48.6
Ours	ViT-B	300	606	164	**49.1**

Table 4: Downstream task evaluation on ADE20K semantic segmentation. We compare our method against other SSL methods and the evaluation settings follow a standard approach.

Method	Backbone	Pretrain Epochs	Object detection			Instance Segmentation		
			AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
Supervised	ViT-B	300	46.9	68.9	51.0	41.5	65.5	44.4
MoCov3	ViT-B	300	45.5	67.1	49.4	40.5	63.7	43.4
DINO	ViT-B	400	46.8	68.6	50.9	41.5	65.3	44.5
BEiT	ViT-B	300	39.5	60.6	43.0	35.9	57.7	38.5
BEiT	ViT-B	800	42.1	63.3	46.0	37.8	60.1	40.6
MAE	ViT-B	300	45.4	66.4	49.6	40.6	63.4	43.7
MAE	ViT-B	1600	48.4	69.4	53.1	42.6	66.1	45.9
CAE	ViT-B	300	48.0	68.7	52.7	42.3	65.6	45.4
SdAE	ViT-B	300	48.9	69.6	53.3	43.0	66.2	46.2
Ours	ViT-B	300	**49.7**	68.1	54.0	**43.1**	65.4	46.7

Table 5: Downstream task evaluation on COCO object detection and instance segmentation. We compare our method against other SSL methods and the evaluation settings follow a standard approach.

size of 1024 with weight decay set to 0.05 and layer decay to 0.65. We compare our model to other SSL methods like DINO[7], BEiT[5], MAE[18] on finetuning. Interestingly we show better performance than MAE [18] with only 300 epochs compared to MAE with 1600 epochs. We also outperform iBoT [31] with out requiring large number of local crops (10) and less epochs of training. We also evaluate on other smaller datasets like Pets, Cars, Cifar10, Cifar100. The finetuning is done for 1000 epochs with a similar settings to DINO[7] in Table 3. When compared against other methods we find that our method performs better thus excelling at both large and smaller datasets.

4.2 Results on semantic segmentation, object detection, instance segmentation

We perform the finetuning and evaluation for semantic segmentation on ADE20K dataset with results provided in Table 4. We follow previous SSL approaches[7, 9, 17] by utlising Upernet[29] as the task layer. We follow MAE[18] for setting the learning rate. Semantic segmentation shows the capability of the network to capture semantic details at various levels and provides for a good downstream task. Our models excels other competing models with less number of pretraining epochs for the complex task of semantic segmentation. Following MAE[18] we finetune our architecture on COCO dataset[21] with results in Table 5. We use Mask R-CNN[16] and we also apply FPN[22] to train the network. The learning rate and the schedule used follows MAE[18]. Our method excels the competing methods in object detection and instance segmentation showcasing its capability to capture mutiple instances.

5 Ablation

We conduct ablation study on Imagenet-1K with ViT as the backbone. We study the usage of multiple teachers for different pretext task. We explore the effect of EMA having different update frequencies i.e. per epoch or per iteration on each of the teacher. We also explore the effect of longer pretraining and different losses in our architecture. Augmentations are generally used in all clustering frameworks[1, 3, 7]. We explore the effect of simple and complex data augmentations in the context of pseudo labelling pretext task. We use ViT-S for all the experiments except for Table 9. The pretraining and finetuning is done of Imagenet-1K for 100 and 50 epochs respectively. For Table 9 we use ViT-B where we pretrain, finetune

Method	In-1K%	Pretraining Time	Memory
SdAE [3]	74.19	2.5 days	11.2G
MAE [18]	-	12.7 days	10G
Single Teacher(Per iteration)	73.5	2.9 days	12G
Single Teacher(Per epoch)	74.3	2.9 days	12G
Disentangled Teacher(default)	74.8	3.2 days	14.5G

Table 6: Comparison of our method with a single-teacher model (using ViT-S). We also include pretraining time and memory usage(ViT-B).

Backbone	Initial Momentum	Final Momentum	In-1K
ViT-S	0.96	1	73.6
ViT-S	0.996	1	74.8

Table 7: Results of using different momentum's for our cluster teacher. Large momentum's work better for generating pseudo labels.

Backbone	λ_m	λ_c	λ_p	In-1K
ViT-S	1	0	0	74.19
ViT-S	1	0	1	73.20
ViT-S	1	1	0	74.23
ViT-S	0	1	1	74.09
ViT-S	1	1	1	74.82

Table 8: Results of our model for different values of scaling factors λ_m, λ_c, λ_p. All the methods are pretrained for 100 epochs and finetuned for 50 epochs.

Data Aug.	Multifold Pseudo Lab.	In-1K
Simple	✗	83.6
Simple	✓	83.8
Complex	✗	83.9
Complex	✓	84.1

Table 9: Table showcasing effect of multifold clustering and data augmentation for pseduo labelling pretext task with ViT-B pretrained for 300 epochs.

for 300, 100 epochs respectively.

Effect of Multiple Teachers and EMA update frequency, schedule We compare a single teacher to disentangled ones (Table 6). Teachers update per epoch or iteration. Our disentangled method, with one teacher for pseudo-labeling and another for reconstruction, outperforms single teachers. Using two teachers allows separate EMA schedules for stable training. We analyze memory and pretraining time for different teachers on ViT-B with 300 epochs and batch size 768, finding dual teachers marginally increase time and memory. Unlike MAE [18], our approach reduces pretraining time with slightly more memory.

We explore momentum schedules for the Pseudo labeling teacher in Table 7. As SdAE [3] studied momentum for the reconstruction teacher, we focus on schedules with pseudo labelling teacher with initial momenta of 0.96 and 0.996, reaching 1. The 0.996 schedule performs better, showing the need for frequent, smaller updates. This highlights the importance of disentangled teachers for different pretext tasks.

Effect of different pseudo label loss terms We study the effects of different auxiliary pretext pseudo labelling loss terms on the performance of the model in Table 8. We find that when we apply only pseudo label loss on patches with i.e. $\lambda_p = 0$, we get a lower accuracy. We find that when we apply both class and patch level pseudo labelling loss it leads to a slight improvement in performance. This proves that auxiliary task of pseudo labelling is required for better performance.

Effect of Multifold clustering We examine the impact of multifold pseudo labeling and data augmentation in Table 9. Multifold masking [10] improves the encoder by maintaining sufficient mutual information, a concept we extend to pseudo labeling. Our results show that multifold pseudo labeling outperforms standard clustering. Additionally, we find that simple augmentation [18] doesn't work well with pseudo labeling, requiring more complex augmentations [7].

6 Conclusion

We introduce an MIM model with an auxiliary pretext task of pseudo label generation to enhance instance discrimination globally and locally. Initially, adding this task alone doesn't yield extra gains due to the distinct nature of pretext tasks. To address this, we propose two disentangled teachers: one for pseudo labeling and another for token-based reconstruction. Token reconstruction requires a stable teacher, updated with Exponential Moving Average (EMA) at each epoch, while pseudo labeling requires more frequent EMA updates at each iteration. We stress the importance of capturing information at multiple levels for better encoder performance. Our future work aims to extend this to multimodality and study teacher behavior for different multimodal tasks.

7 Reviw feedback

Review Response for SNfV: The equation 1 has been fixed with proper changes. Line 186 has been updated to make the reconstruction clear. Similarly the corresponding suggestions have been incorporated. The major problem also seems to be coming from misunderstanding that only N' tokens are being reconstruction by student where the actual reconstruction is being done for $N - N'$ tokens excluding the class token. This is similar to MAE, where after the encoder, mask tokens are introduced in masked regions to make the total tokens same as unmasked input i.e. N. Hence, we mention this in brief in line 187 and adjusted the remaining equations and text.

Review Response for 7viJ: It is prohibitively difficult to conduct experiments with ViT-L due to availability of resources. Similarly, ViT-S is missing from all MAE based methods [10, 11, 18]. Although ablations with ViT-S provided do show that our method works universally. The experiment in Table 6 show that when both teachers use similar momentum update per epoch, iteration, it performs worse. This proves that having different schedules is required for better combination of different tasks. The Table 4, 5 have different metrics because one of the task is semantic segmentation while the other Object detection/Instance segmentation. The metrics are decided by the computer vision community to tailor to individual task and to capture their complexity. We see performance in ViT-S scaling to ViT-B, which led us to use this architecure for ablations. In addition the lack of resources makes it difficult to run all ablations with ViT-B. The results are close in Table.9 but to improve by even 0.1 in Imagenet-1K is a difficult thing. Also the table shows that multifold clustering is better and it requires a complex augmentation applied to teacher.

Review Response for XcH6: L145-147 claim has been validated in Table.6 where our method performs better than MAE with less epochs and shows high performance. It is prohibitively difficult to conduct experiments with ViT-L due to availability of resources. We agree that performing a third experiment with initial momentum set between 0.96 and 0.996 would benefit, but this experiment couldnt be done due to time constraint. Linear probing is not a good metric for MIM based methods as evidenced by [18]. It is prohibitively difficult to conduct experiments with ViT-L due to availability of resources

References

[1] Sara Atito Ali Ahmed, Muhammad Awais, and Josef Kittler. Sit: Self-supervised vision transformer. *ArXiv*, abs/2104.03602, 2021.

[2] Mahmoud Assran, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Michael G. Rabbat, and Nicolas Ballas. Masked siamese networks for label-efficient learning. In *European Conference on Computer Vision*, 2022.

[3] Sara Atito, Muhammad Awais, Ammarah Farooq, Zhenhua Feng, and Josef Kittler. Mc-ssl0. 0: Towards multi-concept self-supervised learning. *arXiv preprint arXiv:2111.15340*, 2021.

[4] Sara Atito, Muhammad Awais, and Josef Kittler. Gmml is all you need. *ArXiv*, abs/2205.14986, 2022.

[5] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ArXiv*, abs/2106.08254, 2021.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *ArXiv*, abs/2006.09882, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[8] Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Mixed autoencoder for self-supervised visual representation learning. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22742–22751, 2023. URL https://api.semanticscholar.org/CorpusID:257834069.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.

[10] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *ArXiv*, abs/2202.03026, 2022. URL https://api.semanticscholar.org/CorpusID:246634394.

[11] Yabo Chen, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, 2022. URL https://api.semanticscholar.org/CorpusID:251223971.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[15] Zhanzhou Feng and Shiliang Zhang. Evolved part masking for self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10386–10395, 2023.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019.

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2021.

[19] Lang Huang, Shan You, Mingkai Zheng, Fei Wang, Chen Qian, and T. Yamasaki. Green hierarchical vision transformer for masked image modeling. *ArXiv*, abs/2205.13515, 2022. URL https://api.semanticscholar.org/CorpusID:249097898.

[20] Xiang Li, Wenhai Wang, Lingfeng Yang, and Jian Yang. Uniform masking: Enabling mae pre-training for pyramid-based vision transformers with locality. *ArXiv*, abs/2205.10063, 2022. URL https://api.semanticscholar.org/CorpusID:248965120.

[21] Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 936–944, 2016. URL https://api.semanticscholar.org/CorpusID:10716717.

[23] Jihao Liu, Xin Huang, Jin Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6252–6261, 2022. URL https://api.semanticscholar.org/CorpusID:257900866.

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[26] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018.

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021.

[28] Chen Wei, Haoqi Fan, Saining Xie, Chaoxia Wu, Alan Loddon Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14648–14658, 2021.

[29] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. *ArXiv*, abs/1807.10221, 2018.

[30] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9643–9653, 2021.

[31] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Loddon Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *ArXiv*, abs/2111.07832, 2021.

8 Acknowledgements:

This work was supported in part by the EPSRC grants MVSE (EP/V0 02856/1) and JADE2 (EP/T022205/1).