MAGMA: Manifold Regularization for MAEs

Alin Dondera*,1, Anuj Singh*,1,2, Hadi Jamali-Rad1,2
[1]Delft University of Technology (TU Delft), The Netherlands
[2]Shell Global Solutions International B.V., Amsterdam, The Netherlands
a.e.dondera@student.tudelft.nl, {a.r.singh, h.jamalirad}@tudelft.nl

Abstract

Masked Autoencoders (MAEs) are an important divide in self-supervised learning (SSL) due to their independence from augmentation techniques for generating positive (and/or negative) pairs as in contrastive frameworks. Their masking and reconstruction strategy also nicely aligns with SSL approaches in natural language processing. Most MAEs are built upon Transformer-based architectures where visual features are not regularized as opposed to their convolutional neural network (CNN) based counterparts, which can potentially hinder their performance. To address this, we introduce MAGMA, a novel batch-wide layer-wise regularization loss applied to representations of different Transformer layers. We demonstrate that by plugging in the proposed regularization loss, one can significantly improve the performance of MAE-based models. We further demonstrate the impact of the proposed loss on optimizing other generic SSL approaches (such as VICReg and SimCLR), broadening the impact of the proposed approach. Our code base can be found here: https://github.com/adondera/magma

1. Introduction

Self-supervised learning has made significant progress over the recent years by producing results on par with supervised baselines [3,5–7,13,34], thus rendering it as a promising paradigm for learning representations without access to labels. Many notable approaches in self-supervised learning such as contrastive learning [7], clustering-based methods [5], redundancy minimization [3, 34] and distillation-based methods [13] aim to learn representations that generalize well by avoiding degenerate solutions and representational collapse by utilising a joint embedding architecture to enforce consistency between representations of different image-views. Inspired by natural language processing (NLP), Masked Autoencoders (MAE) approach the task of self-supervised pre-training by a conceptually simple idea of masking a portion of the input data to then learn to predict the removed content. Specifically, this is applied to images by masking a very large portion (eg. 75%) of their content by replacing it with random patches. This creates a challenging pretext task for image representation learning that requires the neural network to develop a holistic understanding beyond low-level image statistics [14]. By masking a large part of the image and processing only the unmasked region, MAEs provide a computationally efficient way of pre-training large-scale vision transformers such as ViT-B/H/S [12, 14]. However, due to the lack of an objective that optimizes for contrasting negative pairs of images, the features learnt by MAE pre-training require large amounts of labeled data to be fine-tuned for satisfactory downstream task performance [21]. Moreover, deep architectures such as convolutional neural networks are designed with inherent regularization characteristics such as translation invariance, equivariance, and parameter sharing that are relevant to learning information-rich features from images for multiple vision-oriented tasks. On the other hand, ViT-based architectures operate on patches of images and lack these aforementioned regularization characteristics in their feature extraction process. In an ideal scenario, a well-trained network should exhibit a crucial property: if two similar inputs are fed into the network, their resulting outputs should also be close together. This principle ensures that the network learns robust representations that capture the underlying structure of the data, not just random noise or specific details. Deviations from this principle can indicate the network is overly sensitive to small input variations, leading to poor generalization performance on unseen data. One way to enforce this behavior is through manifold regularization, which aims to guide the model toward learning smoother representations aligned with the intrinsic data geometry [4]. To this end, we introduce MAGMA, a novel batch-wide loss that regularizes representations across multiple different layers of a feature extractor. Our extensive experiments, ablations, and analyses empirically demon-

[*] equal contribution

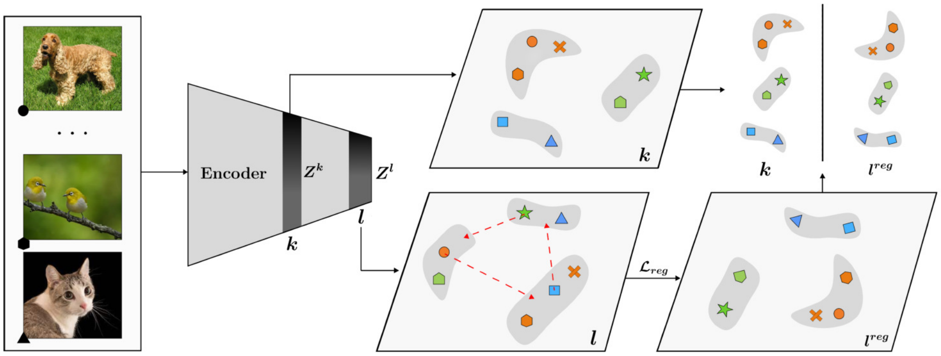


Figure 1. Visualization for the proposed regularization loss MAGMA with MAE: MAGMA penalizes representations that are close in the latent space of intermediate layer k but far apart in layer l latent space. This induces a regularization effect across different layers that preserves inter-sample and intra-batch relationships thus enforcing consistency in the latent representation space. Note that we demonstrate this for MAE based pre-training with a transformer encoder-decoder architecture such as ViT.

strate improved downstream image classification performance on MAE-based baselines by simply plugging in the proposed regularization loss during the pre-training phase. To corroborate the general applicability and broader impact of MAGMA, we demonstrate improved and on-par performance of other generic SSL approaches such as VICReg [3] and SimCLR [7] when pre-trained with our proposed loss.

2. Related Work

Self-supervised learning. Self-supervised learning (SSL) is crucial for overcoming the limitations of traditional supervised learning, which requires vast amounts of expensive, hand-labeled data. By automatically generating labels from the data itself, self-supervised techniques enable models to learn meaningful representations from unlabeled data, reducing our reliance on manual annotation. A common strategy in self-supervised learning (SSL) is to exploit augmentation invariance. By enforcing similarity between augmented views of the same image, these methods aim to learn representations that are robust to common image transformations. To prevent model collapse, various techniques are employed, such as negative sampling [7], cluster assignments, [5], feature decorrelation [3, 34], or asymmetric architectures [13].

Masked Autoencoders. The success of self-supervised learning in Natural Language Processing (NLP), particularly with masked language modeling techniques in models like BERT [11] has inspired analogous developments within computer vision. Masked Autoencoders (MAEs) [14] take the idea of masking and apply it to an autoencoder structure with a pixel-level reconstruction loss. This results in impressive performance across various downstream tasks [8, 24, 35, 37]. Other similar works include BEiT [2], SimMIM [32], and iBOT [36], with close connections to contrastive learning [18, 35].

Manifold regularization. At the core of MAGMA lies the seminal piece of work of [4]. The authors provide a geometrically intuitive and novel semi-supervised learning framework that leverages the underlying geometry of data distributions under the assumption that two points close together on the manifold (i.e., similar in the true underlying structure of the data), should have their corresponding target outputs also be similar. This idea has been successfully applied in deep learning across of variety of tasks, such as speech recognition [28, 29], NLP [22, 33] and vision [15–17, 26], showcasing its usefulness in the general setting. MAGMA extends the concept of manifold regularization to the self-supervised setting, guiding internal network transformations to promote smoother, more generalizable representations. While [26] explores a similar direction, their approach relies on Siamese networks to explicitly calculate similarity measures between input images. In contrast, our regularization operates directly on the representations generated within the network, offering a more tightly integrated self-supervised mechanism.

3. Method: MAGMA

Given an unlabeled dataset $\mathcal{D}_u$ with samples $x \in \mathcal{D}_u$ our goal is to train an encoder f_θ with L layers to produce information-rich representations in a self-supervised fashion. During inference, the parameters of the encoder are frozen θ and a linear layer is trained in a supervised fashion. This procedure is known as linear probing and is the commonly adopted setup in SSL literature. We denote a batch of B samples as $\mathcal{B}$. In this setting, our goal is to apply a batch-level regularization loss in a layer-wise fashion on a set of layers $\mathcal{K} \subseteq [L]$:

$$\mathcal{L}(\mathcal{B}, \mathcal{K}; \theta) = \mathcal{L}_{SSL}(\mathcal{B}; \theta) + \lambda \mathcal{L}_{Reg}(\mathcal{B}, \mathcal{K}; \theta),$$

where the first term denotes a standard self-supervised learning loss, and λ is a weighting parameter between the two terms. While any set of layers can in practice be adopted for such a regularization, we demonstrate later on that applying this on an intermediate and the last layer $\mathcal{K} = \{l, L\}$ would yield the maximum impact. Notably, this is applied only at the pretraining phase.

3.1. Batch-Wide Layer-Wise Manifold Reg.

We denote the representation output of layer $l \in [L]$ of f_θ for input image $i \in \mathcal{B}$ as $Z_i^{(l)}$. Inspired by [4], we propose to apply the following batch-wide layer-wise regularization term to enforce consistency among the output representations of the selected layers:

$$\mathcal{L}_{\text{Reg}}(\mathcal{B}, \mathcal{K}; \theta) = \frac{1}{B^2} \sum_{k,l \in \mathcal{K}} \sum_{i,j \in [B]} w(Z_i^{(k)}, Z_j^{(k)}) \cdot ||Z_i^{(l)} - Z_j^{(l)}||^2 \tag{1}$$

where $w(.) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ can be any similarity kernel, D being the size of the vectorized version of Z. In our study, we employ the Radial Basis Function (RBF) kernel due to its favorable properties as discussed in [4]. Thus, we have:

$$w(Z_i^{(k)}, Z_j^{(k)}) = \exp\left(\frac{-||Z_i^{(k)} - Z_j^{(k)}||^2}{2\sigma}\right)$$

where σ is a free parameter. We choose $\sigma^2 = \texttt{var}(d_{ij})$, with $d_{ij} = ||Z_i^{(k)} - Z_j^{(k)}||^2$, following the approach in [25] for enhanced training stability. Dynamically adjustment of σ this way ensures our regularization adapts to the spread of features inside a batch: the more spread out the features are (i.e. higher σ) the wider the influence of the RBF kernel. Conversely, a lower spread would result in a more focused kernel (focusing on finer, more local distinctions). Note that in Eq. 1, layer k is considered as the reference layer and layer l is regularized accordingly. More concretely, if two instances (Z_i and Z_j) have closer representations in the manifold space of layer k (leading to higher $w(Z_i^{(k)}, Z_j^{(k)})$), but are far apart in the manifold space of layer l, $\mathcal{L}_{\text{Reg}}$ would heavily penalize them, as a result pulling them closer in the regularized manifold. We illustrate later on that in practice these regularizations would not only regularize layer l but also all the previous layers.

The regularization loss in Eq. 1 can be reformulated in terms of the Laplacian matrix L determined by all pairs of instances $(Z_i^{(k)}, Z_j^{(k)})$ in a batch, and is defined as follows:

$$\mathcal{L}_{\text{Reg}}(\mathcal{B}, \mathcal{K}; \theta) = \frac{1}{B^2} \texttt{Trace}(Z^{(l)T} L Z^{(l)}),$$

We make use of the normalized Laplacian for better stability during training, defined as follows:

$$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}},\ D_{ii} = \sum_j W_{ij} = \sum_j w(Z_i^{(k)}, Z_j^{(k)}),$$

Application to Transformers. For the sake of generality, we have so far formulated the problem so that it would be readily applied to any layered neural network architecture. Even though, we have only observed significant impact on ViT based architectures. The only difference for ViT based architectures is that per layer l we would have P patches each of which returning a representation $Z_{i,p}^{(l)}, \forall p \in [P]$, where the image level representation would simply be the average of all those representations $Z_i^{(l)} = \sum_p Z_{i,p}^{(l)}$. The reason behind this averaging strategy is that applying the regularization over the representations of individual patches across different images is not ideal due to patch noise and lack of global context. This may result in irrelevant computations since similar patches within an image already share context through self-attention.

4. Impact of Architectures and Pretraining

The proposed regularization can be seamlessly incorporated into various self-supervised methods, with the caveat that the chosen architecture and pretraining approach play an important role in determining the efficacy of the regularization. The inherent characteristics of CNN-based architectures can diminish the impact of regularization. For instance parameter sharing, translation invariance and equivariance in CNNs, which facilitates the reuse of learned features across various input regions, can result in reduced regularization impact. In contrast, Transformers lack these specific characteristics, potentially making them more suitable for this regularization.

The nature of the pretraining method significantly influences the impact of regularization. Contrastive methods (*e.g.*, SimCLR, MoCo), clustering approaches (SwAV), distillation techniques (DINO, BYOL), and InfoMax/Dimension Contrastive methods all aim to bring representations of augmented views of the same image closer together, essentially performing a task related to our proposed regularization. Therefore, the proposed regularization will have a diminished impact on these methods. On the other hand, Masked Autoencoders (MAE)'s exhibits a generative nature, by randomly masking large portions of an image and reconstructing the missing pixels. Since this process is applied individually, it is also not sharing any information between representations within a batch. These characteristics make it better suited for the regularization term. As a result, our study will primarily focus on MAEs as they align well with the objectives of our proposed regularization approach.

5. Experiments

Our goal in this section is to evaluate the impact of adding our regularization term on top of pre-existing SSL methods, both quantitatively and qualitatively. We aim to address the following questions:

Table 1. Linear probing accuracy and k-nn accuracy (k=10) of models pretrained and evaluated on the given datasets. Adding our proposed regularization term to the baseline method generally increases performance.

	CIFAR-100		STL-10		Tiny-ImageNet		ImageNet-100	
Method	linear	knn	linear	knn	linear	knn	linear	knn
MAE	38.2	36.6	66.5	62.0	17.8	17.7	58.0	47.5
M-MAE (ours)	**43.3**	**40.7**	**71.0**	**65.9**	**20.9**	**20.5**	**69.0**	**49.8**
U-MAE	45.3	45.9	74.9	72.1	21.5	19.0	69.5	56.8
MU-MAE (ours)	**46.4**	**46.4**	**75.6**	**73.0**	**25.2**	**23.9**	**73.4**	**60.1**
SimCLR	62.8	58.7	**90.4**	**86.9**	**50.9**	43.5	67.8	65.3
M-SimCLR (ours)	**63.2**	**59.4**	**90.5**	**86.9**	**51.0**	**44.6**	**68.7**	**65.6**
VICReg	63.6	60.8	**87.4**	**84.5**	45.2	**40.5**	68.4	62.1
M-VICReg (ours)	**64.7**	**61.9**	**87.4**	**84.5**	**45.8**	**40.5**	**70.4**	**65.1**

Table 2. Linear probing (LP) and fine-tuning (FT) accuracy of models pretrained on ImageNet-100.

	CIFAR-100		STL-10		Tiny-ImageNet		ImageNet-100	
Method	LP	FT	LP	FT	LP	FT	LP	FT
MAE	31.5	**76.9**	67.8	82.2	27.8	**63.1**	58.0	79.8
M-MAE (ours)	**51.6**	75.6	**84.8**	**85.5**	**43.1**	**63.1**	**69.0**	**80.6**

Table 3. Linear probing results on different versions of ImageNet for MAE U-MAE with and without regularization.

Method	1%	5%	10%	100%
MAE	0.5	2.7	4.3	50.5
M-MAE (ours)	**1.6**	**8.8**	**13.2**	**51.0**
U-MAE	2.9	15.4	25.6	55.3
MU-MAE (ours)	**4.9**	**20.1**	**29.6**	**57.4**

[Q1] How does $\mathcal{L}_{\text{Reg}}$ influence downstream classification?
[Q2] What is the effect of $\mathcal{L}_{\text{Reg}}$ on the training dynamics?
[Q3] What are the important hyperparameters of the proposed regularization?
[Q4] Is the impact of $\mathcal{L}_{\text{Reg}}$ on representations qualitatively noticeable?

Benchmark Datasets. We evaluate our proposed regularization on commonly adopted datasets for the downstream task of image classification, namely, CIFAR100 [19], STL-10 [9], Tiny-Imagenet [20], and ImageNet-100 [27]. This selection of datasets provides various challenges in terms of data resolution, number of classes, and overall complexity of context presented in the sample image. By testing across diverse datasets, we showcase the robustness and generalizability of our proposed regularization.

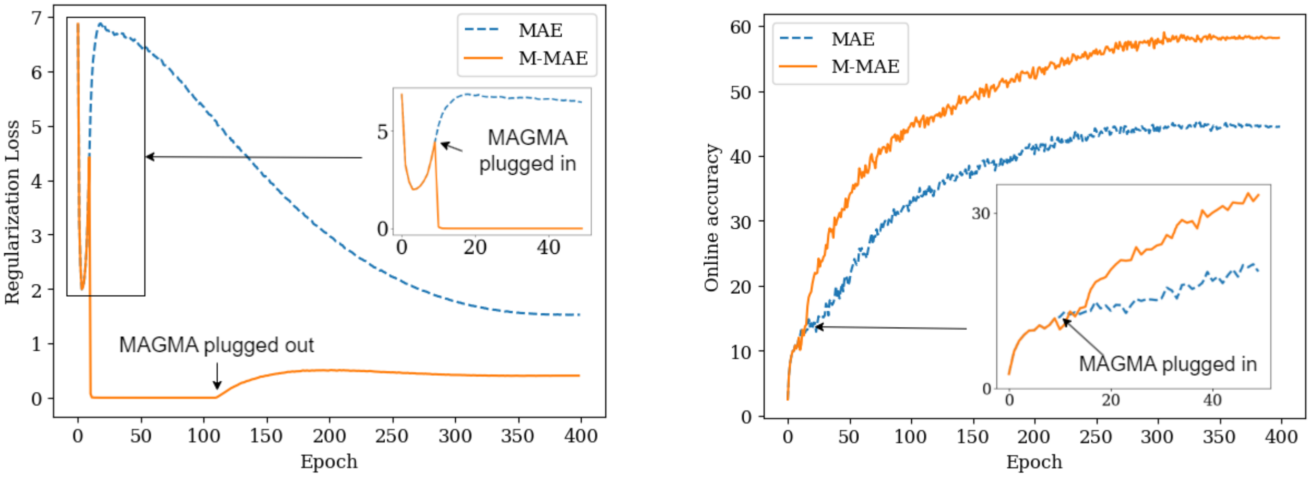
Baseline methods. We evaluate the efficacy of our proposed regularization on several SSL methods (to demonstrate its versatility), with an emphasis on MAE for the reasons discussed in Sec. 4. This includes U-MAE [35], an improvement over the baseline MAE addressing dimensional collapse with an additional regularization term. Additionally, we investigate the impact on two other widely adopted SSL baselines: SimCLR [7] and VICReg [3].

Implementation details. We focus on the impact of regularization by keeping the architectural and hyperparameter choice intact throughout the experimentation, except for the ablation studies. For low(er)-resolution datasets (CIFAR100, STL-10, and Tiny-ImageNet) we use a ViT-Tiny backbone, while for ImageNet-100, we use ViT-Base. We select the best-performing hyperparameter setting for each baseline method and add our regularization on top of it. For our regularization, we tune three parameters: the regularization weight λ, the number of warmup epochs e_{st}, and the duration of the regularizer e_{dur}. More details on the choice of (hyper)parameters can be found in the supplementary material. Notably, for all the other baselines, we present the reproduction results in one optimized pipeline, which in almost all cases leads to performances over the originally reported results in [3, 7, 14, 35].

Evaluation protocol. For our main results, we follow the commonly adopted protocol in SSL, based on freezing the network encoder after the pretraining phase and training a linear layer on top of it in a supervised fashion. For all baselines, we train for 100 epochs using SGD, using a learning rate of 0.1 with decay at steps 60 and 80, and a batch size of 256. In addition, we also evaluate the k-nearest neighbours (kNN) classification accuracy using $k = 10$ and a Euclidean distance measure.

5.1. [AQ1] Comparison Against Other Baselines

Table 1 summarizes the results of applying `MAGMA` on top of the aforementioned four baselines (MAE, U-MAE, SimCLR, and VICReg). We pretrain and evaluate on the same datasets to showcase the robustness of `MAGMA` over various pretraining scenarios. As can be seen, our proposed approach offers significant improvements across all four



(a) Regularization loss tracked throughout pretraining for MAE and M-MAE, on ImageNet-100

(b) Online accuracy tracked throughout the pretraining phase for MAE and M-MAE, on ImageNet-100.

Figure 2. (a) The regularization loss showed for MAE and M-MAE. For MAE we calculate the loss without backpropagating. For M-MAE, we apply the loss after 10 warmup epochs, and take it out after 100 epochs. (b) The online accuracy was obtained by training a linear layer on the representations produced by the encoder throughout pretraining. The accuracy slightly drops for M-MAE when the regularization kicks in but increases at a significantly higher rate compared to MAE.

datasets by outperforming baseline MAE, both in the linear setting ($+5.1\%$ on CIFAR-100, $+4.5\%$ on STL10, $+2.9\%$ on Tiny-ImageNet and $+11\%$ on ImageNet-100), as well for kNN one (**+4.1**% on CIFAR-100, **+3.9**% on STL-10, **+2.8**% on Tiny-ImageNet, and **+2.3**% on ImageNet 100). For U-MAE, while the improvements are still significant, except for Tiny-ImageNet, they are smaller in magnitude. Going beyond MAEs, the results on SimCLR and VICReg, show some marginal improvement opening the door for further investigation and broader impact. To further demonstrate the impact of MAGMA's enhanced self-supervised representation learning at the pretraining phase, we evaluate classification performance of MAE and M-MAE on different datasets by linear-probing and fine-tuning in Table 2. As can be seen, M-MAE significantly outperforms MAE across all datasets when using linear probing for downstream dataset classification, offering improvements up to 20%. Full fine-tuning (especially in low-data regimes) can lead to overfitting to the target dataset [35], completely defeating the purpose and ruling out the impact of our regularization. This is what we also observe in Table 2, where M-MAE offers similar results as compared to the baseline MAE. To assess the effectiveness of MAGMA in lower-data regimes and the commonly adopted large-scale SSL pretraining, we conducted linear probing evaluations on various versions of ImageNet, where we equally sampled 1%, 5%, 10% and the complete 100% split from each class of the original dataset for training. Table 3 demonstrates that both M-MAE and MU-MAE significantly outperform their respective baselines (MAE and U-MAE) across all partial sampling ratios by atleast 1.5% and up to 8.9%, as well as for the full ImageNet-1k dataset by 0.5% and 2.1%. This shows the efficacy of MAGMA in scenarios not only with limited data but also for large-scale pre-training.

5.2. [AQ2] Training Dynamics

Figure 2a illustrates the value of the proposed loss term ($\mathcal{L}_{Reg}$) throughout training epochs. The dashed line illustrates the scenario in which $\mathcal{L}_{Reg}$ is evaluated but not backpropagated. This curve manifest signs of instability (lack of consistency) in the manifold space of representations (for selected layers 11 and 12). The solid curve shows the impact of backpropagating $\mathcal{L}_{Reg}$ (applying MAGMA at epoch 10) where a sudden change of behavior is apparent upon the introduction of $\mathcal{L}_{Reg}$ in the optimization. The fact that the $\mathcal{L}_{Reg}$ drops drastically instead of ascending (dashed line) after being introduced, together with the stability of the loss after removal (at epoch 110), as well as the consistently better online accuracy of M-MAE as seen in Fig. 2b, could potentially suggest that the optimization is now steered in a different direction, leading to an overall significantly better performance. Based on this, we hypothesize that the e_{st} parameter is best set around the point when the $\mathcal{L}_{\text{Reg}}$ loss would start increasing.

Which layers to regularize on? We have run extensive experimentation to effectively select the target layers for applying MAGMA. It turns out regularizing the last layer with respect to the penultimate layer seems to have the maximum impact, in ViT based architecture. In Fig. 3, we demonstrate that choosing $k = 10$ (11-th later in ViT base architecture)

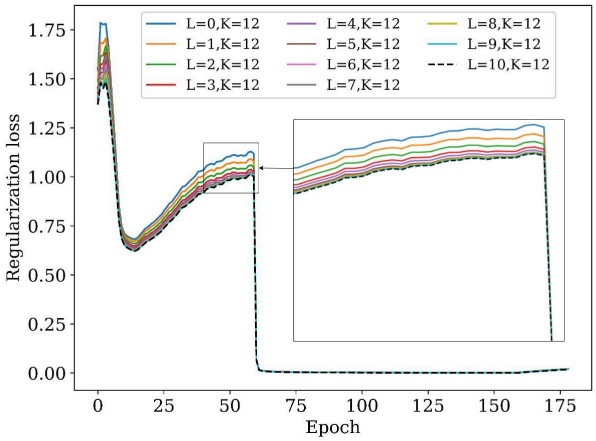


Figure 3. Effect of regularization. Implication: if the representations from any two layers are close, then the output representation will also be close.

as the reference and $l = 11$ (last year) not only leads to regularizing loss across the two layers, but also results in percolated impact through all the previous layers.

5.3. [AQ3] Ablations on Important Parameters

We evaluate the impact of three pivotal regularization parameters (i) λ in Eq. (1), (ii) the epoch at which MAGMAis applied, e_{st}, and (iii) the duration over which the regularization is applied before being plugged out, e_{dur}. The results for the first three parameters are summarized in Tab. 4.

The regularization weight λ directly controls the strength of the regularization effect in the overall optimization loss Eq. (1). Intuitively, lower weight for $\mathcal{L}_{\text{Reg}}$ might not significantly impact the overall optimization, whereas higher weight could lead to an over-regularized optimization and a degraded performance. The results show a similar trend: lowering the weight to 0.1 leads to a performance similar to the baseline ($+2\%$). Increasing the weight by a factor of 10 reduces the gain slightly by 1%. Interestingly, reducing the weight to 0.01 leads to lower downstream classification performance than the baseline. We hypothesize that this is because the regularization introduces a competing gradient signal which inadvertently hinders training performance.

The warm up period e_{st} allows the model to train for a few epochs without $\mathcal{L}_{\text{Reg}}$ to help it establish a reasonable foundation for learning basic representations. This could prevent the regularization from overly restricting the model too early in the training process. As can be seen from Tab. 4, a small number of epochs for e_{st} would already be enough for a maximal impact. Delaying this further seems to have an increasingly negative impact.

Lastly, the duration parameter e_{dur}, determines the amount of pressure put on the model to develop smooth and aligned representations across layers. We experiment with different values ranging from only 10 epochs, up until the end of training (i.e. a duration of 390 epochs). The results show that the impact of this parameter is less pronounced. There is a slight decrease in performance (by about 0.5%), for significantly lower or higher duration periods. It seems that applying MAGMA for a number of epochs already regularizes the representations across the network with a lingering impact from which point onward it can be plugged out without hampering the overall performance. As discussed earlier in Sec. 5.2, we hypothesize that this lingering impact is related to the adjusted optimization landscape as a result of applying the proposed regularization.

To further investigate the sensitivity of MAGMA, we evaluate the performance by changing the backbone architecture starting from small to larger (ViT-S to ViT-L). As can be seen in Tab. 5, increasing the capacity of the backbone results in considerable performance improvement in the baseline approaches (MAE and U-MAE) where the performance boosts decreases for changing the backbone from ViT-B to ViT-L. Interestingly, similarly significant boost can be observed on the MAGMA optimized baselines (M-MAE, MU-MAE), offering consistent improvement over the baselines.

5.4. [AQ4] Qualitative Analysis

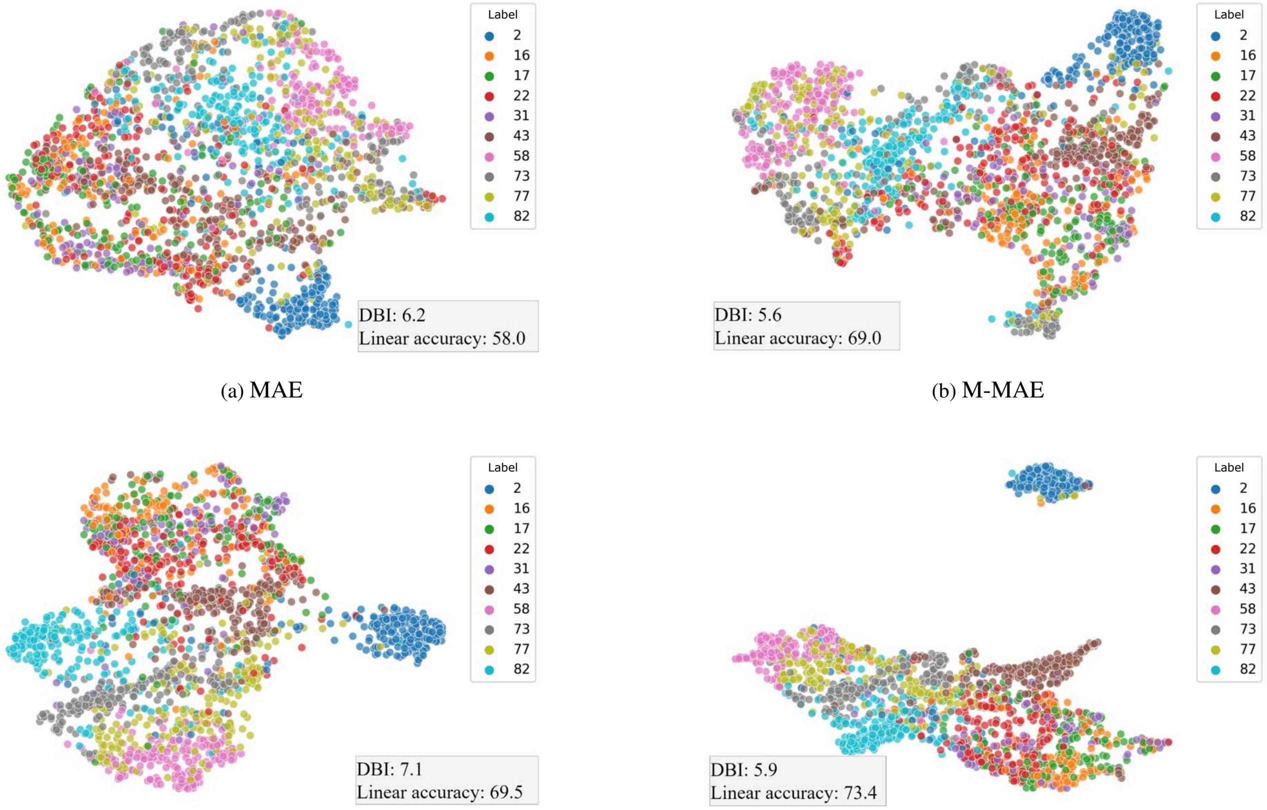
PacMAP. To qualitatively assess the impact of our regularization, we visualize the representations of MAE, U-MAE, as well as their regularized version, M-MAE, and MU-MAE, on a random sample of 10 classes from ImageNet-100. We use PacMAP [31] for dimensionality reduction. PaCMAP outperforms t-SNE [30] and UMAP [23] in preserving the global structure of high-dimensional data within visualizations. This means it more accurately reflects the large-scale relationships and patterns present in the original dataset. We include the linear accuracy, as well as the Davies-Bouldin Index (DBI) [10] alongside the visualizations. DBI is a common metric used to evaluate clustering algorithms, where lower DBI scores indicate better separation between clusters and tighter groupings within clusters.

Results are shown in Fig. 4. Comparing (a) MAE with (b) M-MAE, we observe a slight improvement in the clustering structure after applying MAGMA. The M-MAE representations exhibit tighter clusters with better class separation, leading to a higher linear accuracy (69.0 vs. 58.0) and a lower DBI (5.6 vs. 6.2). Similarly, comparing (c) U-MAE and (d) MU-MAE reveals that incorporating MAGMA into U-MAE further refines the representation space. While U-MAE already improves upon the baseline, MU-MAE achieves even tighter clusters and greater inter-class separation, resulting in a further boost in accuracy (73.4) and a lower DBI (5.9). This highlights the complementary nature of MAGMA and U-MAE, where MAGMA enhances the already improved representations learned by U-MAE.

PCA. Inspired by [1], we take our pretrained MAEs

Table 4. Linear accuracy performance using different choices of hyperparameters for regularization. Results computed on ImageNet-100.

λ	*1*	*0.1*	*0.01*	*10*	1	1	1	1	1	1	1	1
e_{st}	10	10	10	10	*0*	*2*	*20*	*50*	10	10	10	10
e_{dur}	100	100	100	100	100	100	100	100	*10*	*50*	*200*	*390*
Accuracy	69.0	60.0	54.6	67.9	68.2	69.0	65.5	62.7	68.5	68.8	69.0	68.5



(a) MAE

(b) M-MAE

(c) U-MAE

(d) MU-MAE

Figure 4. PaCMAP plots for MAE-based methods. Applying `MAGMA` on top of U-MAE leads to compact and well-defined clusters.

Table 5. Linear probing with different ViTs on ImageNet-100.

Method	ViT-S	ViT-B	ViT-L
MAE	46.8	57.9	60.6
M-MAE (ours)	**61.2**	**69.2**	**73.9**
U-MAE	57.6	69.5	**78.2**
MU-MAE (ours)	**62**	**73.4**	**78.4**

and extract features from each patch, and each layer (in this case, we isolate the *key* features from the self-attention mechanism), apply PCA, and take the leading component. We use upsampling to obtain a heatmap of the same resolution as the original image. This provides a qualitative analysis of the quality of the intermediate representations learned by the models, showcasing the impact of the added regularization term. One visible pattern is the reduction in noise, specifically in the first and last layers, that M-MAE exhibits when compared to its MAE counterpart.

Attention maps. We investigate the impact of the different regularizations on the self-attention maps of the ViT-B architecture's last layer. To this end, we randomly select images from the ImageNet-1K validation set and visualize their corresponding attention maps in Fig. 6. Our observations reveal that the baseline MAE model often tends to attend to the background of the image, in line with findings from prior work [21]. In contrast, we notice differences when applying `MAGMA`: it appears to promote a semantic separation of the attention focus, where the model

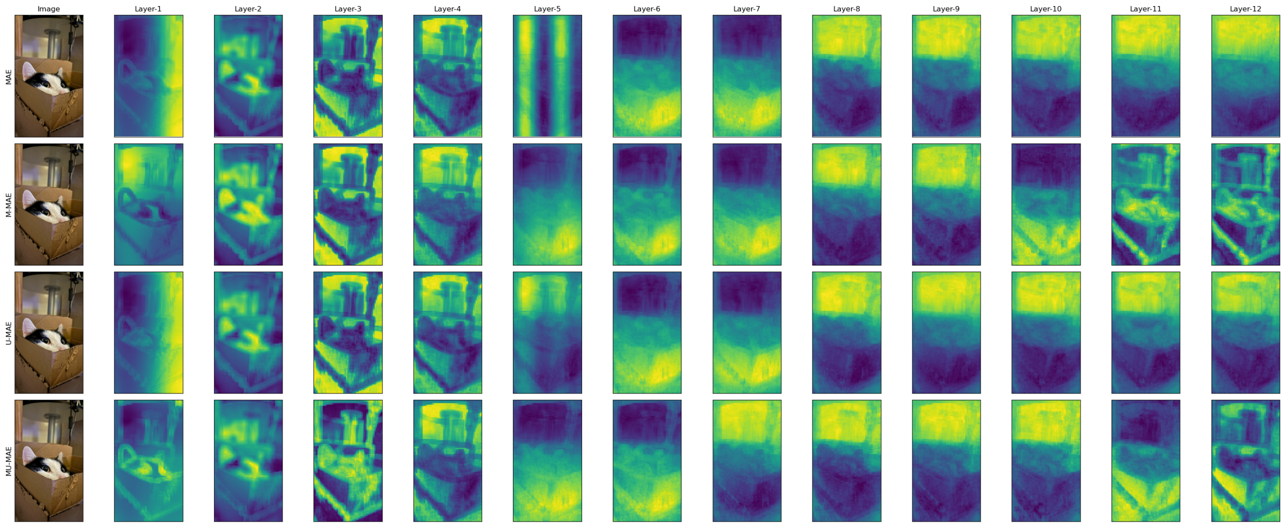


Figure 5. Visualization of PCA's leading component for features extracted from different layers of a ViT-B pretrained using MAE, M-MAE (ours), U-MAE, and MU-MAE (ours).

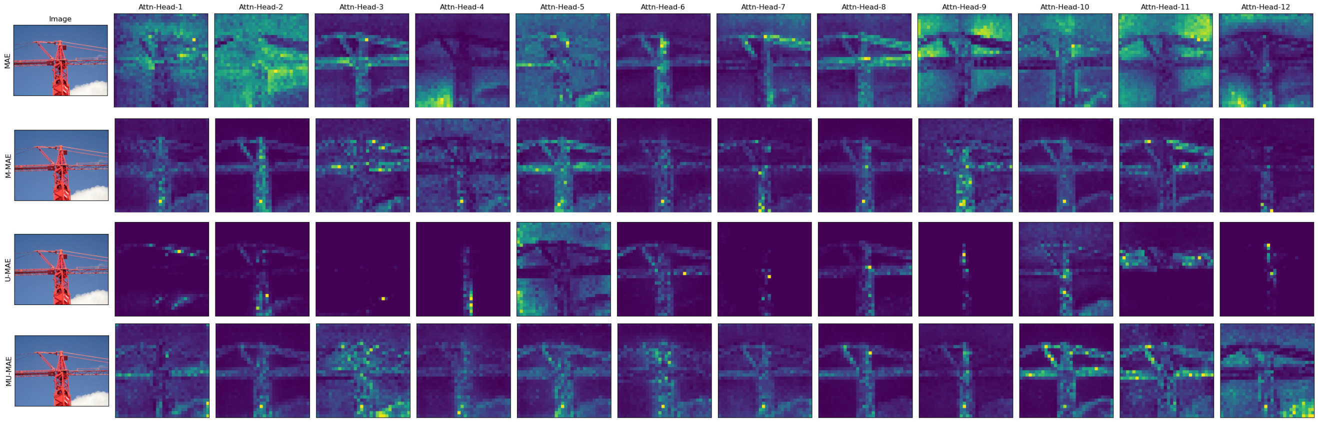


Figure 6. Attention maps from the 12 attention heads of the last layer of a ViT-B. The maps are extracted over the four MAE-based methods evaluated: MAE, M-MAE (ours), U-MAE, MU-MAE (ours)

tends to attend primarily to either the background or the central object, but rarely both simultaneously. This suggests that MAGMA guides the model towards learning more specialized and semantically coherent representations, improving its ability to distinguish between foreground and background elements.

6. Concluding Remarks

We propose MAGMA, a novel regularization technique that regularizes the representations and enforces consistency across layers of a transformer-based MAE. We demonstrate the efficacy of the proposed approach through a suite of experimentation resulting in significant performance gain over MAE-based baselines in most scenarios.

Computational complexity. M-MAE offers a 1.5% and MU-MAE a 2% drop in throughput (100MB GPU-memory) compared to their respective MAE and U-MAE baselines, thus adding an insignificant extra computation cost to the baseline methods while keeping the parameter count same all across. More details are included in Section 3 of the supplementary material.

Broader impact. MAGMA can be rather straightforwardly applied to any kind of SSL approach irrespective of the backbone architecture. As discussed in Section 3 this applies to any layered deep networks, irrespective of an encoder-decoder architecture. This potentially broadens the application of MAGMA to contexts even beyond computer vision. This is an avenue for future work.

Limitations. Going beyond ViT-based architectures to CNNs, we observed that the impact of MAGMA is considerably smaller. We argue that standard operations in modern CNN-based architectures (such as average pooling, weight sharing, etc.) might already serve as a regularizer, minimizing the impact of MAGMA. Note that CNN architectures are outside the scope of this work (as reflected throughout the paper), thus, will investigate this in our future work.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 6

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

[3] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021. 1, 2, 4

[4] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006. 1, 2, 3

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 1, 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 4

[8] Zekai Chen, Devansh Agarwal, Kshitij Aggarwal, Wiem Safta, Mariann Micsinai Balan, and Kevin Brown. Masked image modeling advances 3d medical image analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1970–1980, 2023. 2

[9] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011. 4

[10] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979. 6

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

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

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 4

[15] Hongwei Hu, Bo Ma, Jianbing Shen, Hanqiu Sun, Ling Shao, and Fatih Porikli. Robust object tracking using manifold regularized convolutional neural networks. *IEEE Transactions on Multimedia*, 21(2):510–521, 2018. 2

[16] Biao Jie, Daoqiang Zhang, Bo Cheng, Dinggang Shen, and Alzheimer's Disease Neuroimaging Initiative. Manifold regularized multitask feature learning for multimodality disease classification. *Human brain mapping*, 36(2):489–507, 2015. 2

[17] Charles Jin and Martin Rinard. Manifold regularization for locally stable deep neural networks. *arXiv preprint arXiv:2003.04286*, 2020. 2

[18] Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. A mutual information maximization perspective of language representation learning. *arXiv preprint arXiv:1910.08350*, 2019. 2

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[20] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 4

[21] Johannes Lehner, Benedikt Alkin, Andreas Fürst, Elisabeth Rumetshofer, Lukas Miklautz, and Sepp Hochreiter. Contrastive tuning: A little help to make masked autoencoders forget. *arXiv preprint arXiv:2304.10520*, 2023. 1, 7

[22] Ximing Li, Yang Wang, Jihong Ouyang, and Meng Wang. Topic extraction from extremely short texts with variational manifold regularization. *Machine Learning*, 110:1029–1066, 2021. 2

[23] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 6

[24] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022. 2

[25] Pau Rodríguez, Issam Laradji, Alexandre Drouin, and Alexandre Lacoste. Embedding propagation: Smoother manifold for few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 121–138. Springer, 2020. 3

[26] Uri Shaham, Kelly Stanton, Henry Li, Boaz Nadler, Ronen Basri, and Yuval Kluger. Spectralnet: Spectral clustering using deep neural networks. *arXiv preprint arXiv:1801.01587*, 2018. 2

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 4

[28] Vikrant Singh Tomar and Richard C Rose. Manifold regularized deep neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014. 2

[29] Vikrant Singh Tomar and Richard C Rose. Graph based manifold regularized deep neural networks for automatic speech recognition. *arXiv preprint arXiv:1606.05925*, 2016. 2

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[31] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *The Journal of Machine Learning Research*, 22(1):9129–9201, 2021. 6

[32] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 2

[33] Chu Yonghe, Hongfei Lin, Liang Yang, Yufeng Diao, Shaowu Zhang, and Fan Xiaochao. Refining word reesprentations by manifold learning. In *Proc. 28th Int. Joint Conf. Artif. Intell*, pages 5394–5400, 2019. 2

[34] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1, 2

[35] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems*, 35:27127–27139, 2022. 2, 4, 5

[36] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 2

[37] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image analysis. *arXiv preprint arXiv:2203.05573*, 1(3), 2022. 2