I.J. Image, Graphics and Signal Processing, **2023, 5, 29-38**
Published Online on October 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2023.05.03

Denoising Self-Distillation Masked Autoencoder for Self-Supervised Learning

Jiashu Xu*
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 03056, Ukraine
E-mail: xu.jiashu@lll.kpi.ua
ORCID iD: https://orcid.org/0000-0001-6300-3629
*Corresponding Author

Sergii Stirenko
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, 03056, Ukraine
E-mail: stirenko@comsys.kpi.ua
ORCID iD: https://orcid.org/0000-0001-5478-0450

Received: 12 July 2023; Revised: 27 August 2023; Accepted: 16 September 2023; Published: 08 October 2023

Abstract: Self-supervised learning has emerged as an effective paradigm for learning universal feature representations from vast amounts of unlabeled data. It's remarkable success in recent years has been demonstrated in both natural language processing and computer vision domains. Serving as a cornerstone of the development of large-scale models, self-supervised learning has propelled the advancement of machine intelligence to new heights. In this paper, we draw inspiration from Siamese Networks and Masked Autoencoders to propose a denoising self-distilling Masked Autoencoder model for Self-supervised learning. The model is composed of a Masked Autoencoder and a teacher network, which work together to restore input image blocks corrupted by random Gaussian noise. Our objective function incorporates both pixel-level loss and high-level feature loss, allowing the model to extract complex semantic features. We evaluated our proposed method on three benchmark datasets, namely Cifar-10, Cifar-100, and STL-10, and compared it with classical self-supervised learning techniques. The experimental results demonstrate that our pre-trained model achieves a slightly superior fine-tuning performance on the STL-10 dataset, surpassing MAE by 0.1%. Overall, our method yields comparable experimental results when compared to other masked image modeling methods. The rationale behind our designed architecture is validated through ablation experiments. Our proposed method can serve as a complementary technique within the existing series of self-supervised learning approaches for masked image modeling, with the potential to be applied to larger datasets.

Index Terms: Self-supervised learning, Masked Autoencoder, Siamese Networks, Computer Vision

1. Introduction

In recent years, self-supervised learning has gained significant attention in the field of deep learning and has even been referred to as the "dark matter of intelligence" [1]. The development of self-supervised learning has driven the era of large-scale models in deep learning. For instance, a series of natural language models such as ChatGPT [2] and large-scale vision models such as SAM [3] and DINOv2 [4] proposed by META have emerged. These models heavily rely on the support of self-supervised learning algorithms. Self-supervised learning algorithms enable the direct pre-training of a generic model with feature extraction capability using unlabeled raw data. The pre-trained model can then be fine-tuned with a small labeled dataset in downstream tasks.

In the field of computer vision, the self-supervised learning paradigm based on contrastive learning has recently gained significant attention [5-8]. These methods have demonstrated remarkable results in many downstream computer vision tasks. However, with the emergence of transformer [9], a new line of research has been inspired by the masked language modeling task in the NLP field, leading to the development of many self-supervised learning methods based on masked image modeling (MIM) [10-13]. MIM involves reconstructing the masked image patches through visible image patches. The MAE [13] utilizes an asymmetric encoder-decoder architecture to reconstruct the original pixels using masked image patches. MaskFeat [14] reconstructs the Oriented Gradient Histogram (HOG) of the original image. Although MAE has successfully pre-trained large-scale models and demonstrated state-of-the-art performance when fine-tuned on the ImageNet dataset compared to classical contrastive learning experiments, the objective function of

MIM only models the MSE loss at the pixel level of image details. This limitation may result in a failure to learn abstract semantic information, which is a crucial component of image understanding. Consequently, the representations obtained from MIM often require fine-tuning for specific recognition tasks, which may lead to overfitting when the training data is limited. To address this issue, we propose a denoising self-distilling Masked Autoencoder model, which simultaneously considers feature-level and pixel-level losses.

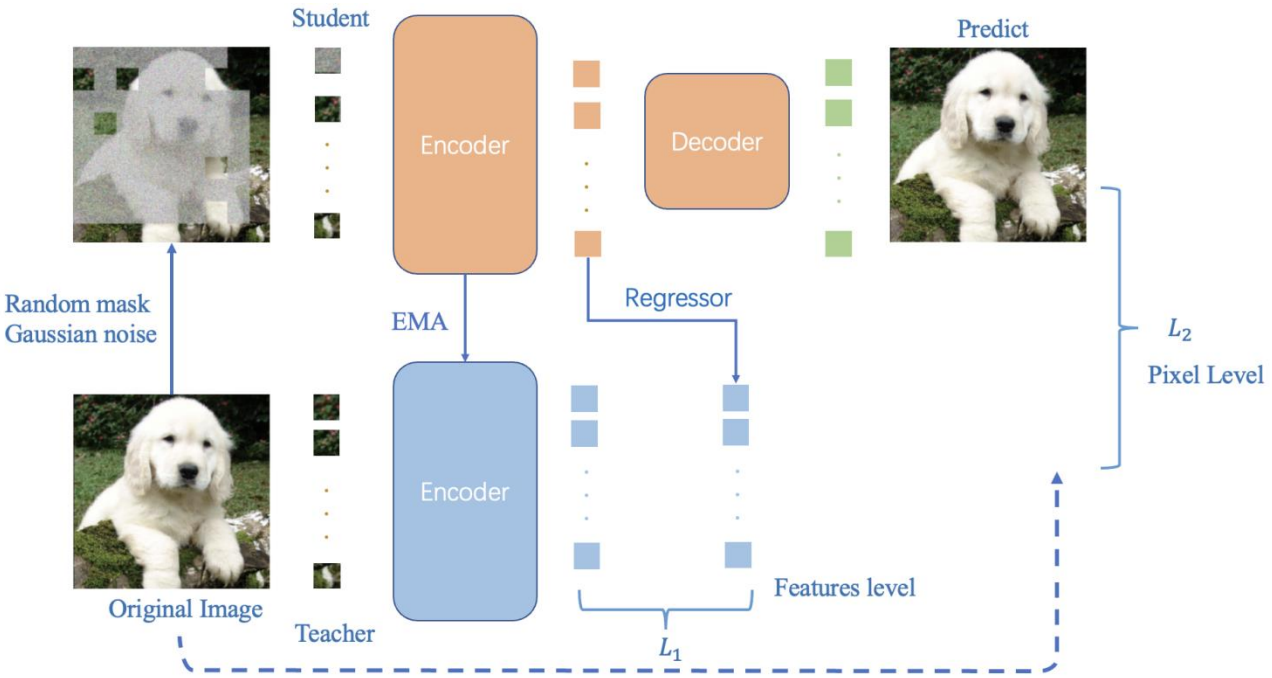

Fig.1. A pipeline for denoising autoencoder with distillation to remove noise from images and map the encoded features to the target features. The decoder component of the autoencoder is composed of two parts: one is responsible for denoising the noisy input to reconstruct the original image, while the other is responsible for mapping the encoded features to the target features. L_1 represents a loss function based on the cosine similarity of features, whereas L_2 is a mean squared error loss function at the pixel level.

The method we propose, as illustrated in Fig.1., aims to eliminate noise from images and map the encoded features to the target features. The teacher network updates its parameters by means of an exponential moving average (EMA) approach, as opposed to gradient-based updates. The student network comprises an encoder, a regressor, and a decoder. The encoder extracts features from the input, which consists of randomly inserted noisy patches. The regressor predicts the target patch features based on the features of the noisy patches. Finally, the decoder maps the predicted noisy patch features back to the original image.

Our contribution can be summarized as follows.

- We propose a denoising self-distillation masked autoencoder architecture as a self-supervised learning method, which utilizes a denoising autoencoder with distillation capabilities to remove noise from images and map the encoded features to target features. By constructing feature-level and pixel-level loss functions, this approach can effectively learn from unlabeled data and extract complex features.
- Our proposed method outperforms classical contrastive learning pretraining on Cifar-10 [15], Cifar-100, and STL-10 [16], and achieves results comparable to other MIM methods. This demonstrates the effectiveness of our proposed method.
- Our ablation experiments analyze the experimental results from different perspectives, including model architecture, noise masking ratio, and the impact of hyperparameter λ in the Eq. 3.

Overall, while computational resource constraints limit our ability to validate our method on large datasets, our experimental results are still sufficient to demonstrate its effectiveness. Our approach is a valuable addition to the field of self-supervised learning in image processing, offering new methodological insights and techniques.

2. Related Work

A. Masked Image Modeling

The Masked Image Model (MIM) draws inspiration from self-supervised learning methodologies that leverage the Masked Language Model (MLM) [17]. The MIM is specifically designed to extract meaningful features from images

that have been either masked or corrupted, thereby enabling the model to learn from unlabeled data. Common approaches in this field utilize the Vision Transformer [9] to model masked images. For instance, iGPT [18] predicts the bottom half of an image using the top half as input. BEiT [10] maximizes the similarity between predicted visual tokens and visual tokens that correspond to real patches, while iBOT [19] proposes an online tokenizer that can learn the MIM objective end-to-end without requiring a separate tokenizer training phase. Moreover, MAE [13] introduced an asymmetric masked autoencoder for reconstructing image pixels, while MaskFeat [14] and Mixup feature [41] are based on this method to reconstruct feature maps. Our model is informed by prior techniques and investigates two aspects: model architecture and feature alignment.

B. Siamese Networks

In the field of computer vision, the combination of Siamese networks and self-supervised learning aims to generate similar image embeddings for two views of an image, enabling the pre-training of an encoder without the need for image annotation [7,20]. However, Siamese networks face the challenge of preventing model collapse while maintaining image augmentation to learn useful representations. Several research works have been dedicated to addressing this issue [20,21,22], leading to a series of developments in utilizing Siamese networks for self-supervised learning [19,26]. The Masked Siamese Networks [23] method selects randomly masked anchor views of images as input to the encoder, aligning the image features with the target view. Additionally, Masked Siamese ConvNets [24] is a self-supervised masking technique applied to convolutional neural networks. SdAE [25] consists of two parts, a teacher network that generates potential representations of masked tags, and a student network that adopts an encoder-decoder structure for reconstructing masked information. Our network structure shares some similarities with the SdAE model, as we have employed the Siamese Network architecture at the feature level.

C. Self-supervised learning

Self-supervised learning has gained significant attention in the field of computer vision, with the objective of designing diverse pretext tasks for pre-training [27,28,29,30]. One important research focus of pretext tasks involves removing parts of the input and learning to reconstruct the missing content. For example, cross-channel prediction [31] and image coloring tasks [30] have been explored. The introduction of context encoders [33] was groundbreaking as it proposed a regression learning task to generate missing image patches based on the surrounding context. Following the emergence of ViT [9], numerous studies have revisited image reconstruction tasks and investigated the utilization of masked autoencoders for model pre-training [13,14,33]. These approaches involve predicting masked image regions at the pixel level or utilizing tokenizers. In contrast, our method not only predicts missing values in the input at the pixel level but also ensures that the global representation of the noisy input aligns with that of the intact input, enabling effective denoising.

3. Method

Our denoising self-distillation masked autoencoder is pre-trained by addressing the task of denoising masked noisy images, and the encoder is obtained through pre-training. The architecture, as illustrated in Fig.1., the architecture comprises encoders, a decoder, and a regressor. In the student network, pixel-level restoration is implemented, making the model focus on local information. The role of the regressor is to map the latent representation of the student network encoder, to the latent representation of the teacher network, making the model focus on global features. Compared to mere pixel reconstruction pretext tasks, our model not only focuses on local information but also on global information.

A. Framework

In each iteration of the pre-training process, we sample a mini batch B of images. For an index $i \in B$, let X_i represent the $i-th$ image in the mini-batch. Each image X_i is split into a set of patches X_{io} using a fixed patch size, and Gaussian noise is applied to randomly selected image patches based on a masking ratio r, resulting in X_{in}. X_{io} is utilized as the input for the teacher network, while X_{in} serves as the input for the student network. The pretext task involves predicting the original patches from the noisy patches, effectively performing denoising.

Encoder

The role of the Student Encoder $f_{\theta_s}(X_{in})$ is to map the noisy patches X_{in} to the latent representation Z_n. This operation encompasses all the patches of the image X_i. The encoder employs the ViT architecture, initiating with patch embeddings and incorporating positional embeddings to preserve spatial information. Subsequently, the combined embeddings undergo processing via a Transformer encoder, ultimately resulting in the generation of Z_n. The parameters of the $f_{\theta_s}(X_{in})$ are updated through gradient-based optimization. Similarly, the Teacher Encoder $f_{\theta_t}(X_{io})$ and the Student Encoder $f_{\theta_s}(X_{in})$ share a similar network structure to accomplish the mapping from X_{io} to the latent representation Z_o. The parameters of the $f_{\theta_t}(X_{io})$ are updated by leveraging an exponential moving average [34] of the $f_{\theta_s}(X_{in})$ parameters.

Decoder

The decoder maps the latent representation Z_n to the denoised patches $Y_n = \phi(Z_n)$. Like the encoder, the decoder also employs the VIT architecture, but they exhibit asymmetric structures. The decoder requires only a few layers of

VIT, resulting in significantly fewer parameters compared to the encoder. However, the decoder only takes the latent representation of the noisy patches and positional embeddings of the noisy patches as inputs.

Regressor

This regressor leverages the latent representation $Z_n{}'$ outputted by the student encoder to predict the latent representation $Z_o{}'$ of the target patches. $Z_n{}'$ and $Z_o{}'$ does not include the latent representation of patches without added Gaussian noise. For feature alignment, we adopt the VIT structure without a linear header to implement the regressor.

B. Objective Function

The objective function of our method comprises two components: the L_2 distance between the predicted denoising patches Y_{in} and the original image patches X_{io}, and the feature cosine similarity between the output of the regression decoder $\phi_r(Z_n{}')$ and the output of the teacher encoder Z_o.

$$L_y = min\frac{1}{N}\sum_{i=1}^{N} D\left(\phi\left(f_{\theta_s}(X_{in})\right), X_{io}\right) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in P_i} \| Y_{in_j} - X_{io_j} \|_2^2 \tag{1}$$

P in Eq. 1 represents the patches set.

$$L_z = min\ log q_{\phi_r}(Z_o'|Z_n') \approx \min\left[1 - \frac{\sum_{i=1}^{n}\phi_r\left(f_{\theta_s}(X_{in})\right) sg[f_{\theta_t}(X_{io})m^i]}{\sqrt{\sum_{i=1}^{n} m^i\left(\phi_r\left(f_{\theta_s}(X_{in})\right)\right)^2}\sqrt{\sum_{i=1}^{n} m^i\left(sg\left[f_{\theta_t}(X_{io})\right]\right)^2}}\right] \tag{2}$$

In Eq. 2, sg[·] represents the stop gradient operation, m^i represents the index of the noise patches.
The final loss of the model is the weighted sum of two losses, as defined by Eq. 1 and Eq. 2.

$$L = \lambda L_y + (1-\lambda)L_z,\ \lambda \in (0,1) \tag{3}$$

The value of λ will be discussed in the experimental section.

C. Distillation Strategy

The parameters of the student network θ_s are updated through backpropagation to minimize the final loss function Eq. 3. On the other hand, the teacher network is updated using an exponential moving average (EMA) in a momentum update manner. Specifically, we update the parameters θ_t of the teacher encoder using the parameters θ_s of the student encoder, as illustrated by the following Eq. 4.

$$\theta_t = \eta\theta_s + (1-\eta)\theta_t \tag{4}$$

Here $\eta \in [0,1)$ is the momentum hyperparameter, which is generally set at 0.99. We set the cosine scheduler to update η

4. Experiments

In this section, we evaluate the feature extraction capabilities of our pre-trained encoder on the CIFAR-10 [15], CIFAR-100 [15], and STL-10 [16] datasets. Specifically, we assess the classification performance of the pre-trained encoder through fine-tuning and linear probing. We then compare our method with other approaches. Finally, we conduct ablation studies on the key components of the proposed method.

A. Implementation Details

We investigated different ViT architectures, namely ViT-tiny (consisting of 12 transformer blocks with a Hidden size of 192), ViT-small (composed of 12 transformer blocks with a Hidden size of 384), and ViT-base (comprised of 12 transformer blocks with a Hidden size of 768). The regression decoder consists of two transformer blocks based on self-attention, while the decoder comprises four transformer blocks based on self-attention and an additional linear projection for prediction.

Table 1. Pre-training settings

config	CIFAR-10 or 100	STL-10
architecture	ViT-T\ViT-S	ViT-S\ViT-B
batch size	2048\2048	2048\1024
patch size	2	6
optimizer	AdamW [35]	
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$	
weight decay	0.05	
learning rate schedule	Cosine decay [36]	
base learning rate	1.5e-4	
warmup epochs [37]	20	
Epoch	500	500
EMA η	cosine scheduler (0.96,0.99)	

Due to the image size of 32x32 in the CIFAR-10 and CIFAR-100 datasets, we partitioned the images into 16x16 patches with a patch size of 2x2 for validation on ViT-Tiny and ViT-Small architectures. Similarly, for the STL-10 dataset with image size of 96x96, we divided the images into 16x16 patches with a patch size of 6x6 for validation on ViT-Small and ViT-Base architectures. The pre-training stage configuration followed the MAE [13] while conducting these experiments. The details of the pre-training settings are shown in Table 1. The visual denoising results of the STL-10 experiment are shown in Fig.2. Owing to the limitations of computational resources, our experiments were conducted solely on a comparatively smaller dataset, our approach is also suitable for larger datasets, and on larger datasets, a similar rule for patch division can be implemented.

B. Evaluation of Pre-trained Model

Evaluation of the Pre-training Feature Extraction Capability through Linear Probing, a commonly adopted approach in self-supervised representation learning, has been widely utilized by numerous researchers [1]. Linear probing involves a supervised process wherein a linear classifier is added on top of the feature outputs from a pre-trained encoder. The parameters of the pre-trained encoder are frozen, and the linear classifier is trained and tested on a validation set to assess the performance.

Different from Linear Probing, full fine-tuning is an alternative approach to evaluate pre-trained models by performing fine-tuning on the entire model. All parameters of the pre-trained model, including the pre-trained encoder, can be adjusted during this process. While Linear Probing primarily focuses on evaluating the feature extraction capability of pre-trained models, fine-tuning is employed to utilize pre-trained models for specific tasks and further improve performance.

Table 2. Self-supervised pre-training methods based on the VIT backbone were evaluated on three datasets, where (a) denotes the fine-tuning experimental results and (b) denotes the linear probing experimental results.

(a) full fine-tuning

Method	pre-trained Epochs	CIFAR-10		CIFAR-100		STL-10	
		ViT-tiny	ViT-small	ViT-tiny	ViT-small	ViT-small	ViT-base
scratch baseline	-	73.88	79.86	51.55	56.17	77.98	82.41
BEiT [10]	500	88.93	90.65	66.32	66.93	84.32	86.22
MoCo v3 [38]	500	88.91	90.88	66.17	67.39	84.61	87.07
MAE [13]	500	88.77	90.26	65.93	66.51	85.63	86.38
MAE [13]	1200	89.87	91.79	66.72	67.83	86.20	87.69
Mask feat [14]	1200	**90.12**	91.75	66.83	**67.86**	86.23	87.85
CAE [11]	300	89.93	91.56	66.84	67.83	86.08	87.76
SdAE [25]	300	89.98	91.83	66.96	67.79	85.90	87.71
Ours	500	89.76	**91.94**	**67.23**	67.77	**86.31**	**87.86**

(b) Linear probing

Method	pre-trained Epochs	CIFAR-10		CIFAR-100		STL-10	
		ViT-tiny	ViT-small	ViT-tiny	ViT-small	ViT-small	ViT-base
BEiT [10]	500	47.68	56.73	27.89	33.78	42.21	48.57
MoCo v3 [38]	500	**76.20**	**77.91**	**50.46**	**54.41**	**78.87**	82.75
MAE [13]	500	73.77	76.78	48.22	51.89	76.10	80.16
MAE [13]	1200	75.87	77.53	50.19	54.08	77.96	82.36
Mask feat [14]	1200	75.58	77.43	50.25	**54.41**	78.32	82.61
CAE [11]	300	74.90	77.15	50.21	54.26	78.48	81.62
SdAE [25]	300	75.17	76.86	50.01	53.71	78.30	82.11
Ours	500	75.53	76.98	50.17	53.84	78.56	**82.79**

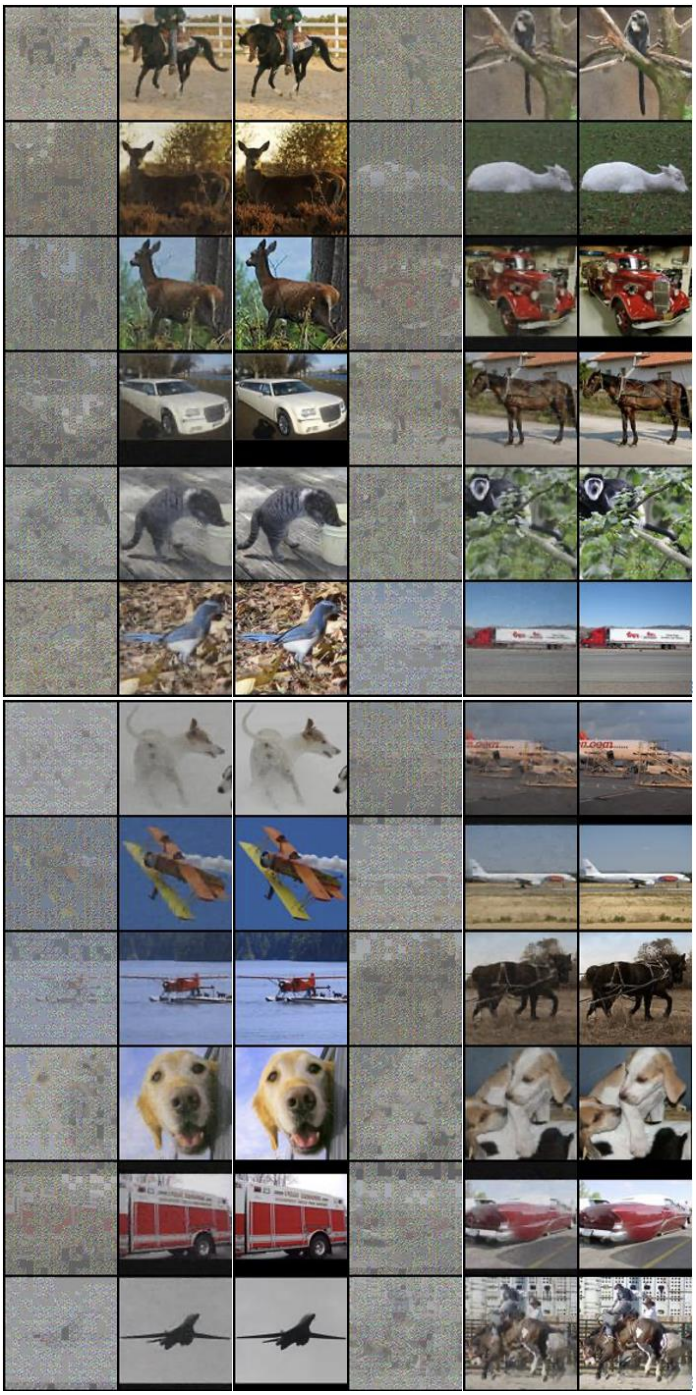

Fig. 2. The visual denoising results of the STL-10 experiment. The first column displays the noisy image, the second column shows the denoised result, and the third column represents the original image. The subsequent three columns correspond to the preceding order.

In the experimental phase of this study, to ensure rigorous comparability and fairness, a selection of seminal works from prior research was employed as benchmarks for comparative analysis. To maintain consistency in the backbone network utilized during experimentation, the self-supervised learning method Moco v3, designed on the foundation of the VIT backbone network, was extracted from the contrastive learning paradigm. Concurrently, from the Masked image modeling domain, representative cornerstone works, including BEIT, MAE, and Mask Feat, were selected. Notably, both CAE and SdAE adopt a bifurcated architecture, demonstrating a predilection for model optimization within the feature domain. Specifically, the student branch of SdAE employs an encoder-decoder framework, reconstructing omitted information from the input imagery, while its teacher branch is dedicated to generating latent representations of masked tokens. Our methodology uniquely optimizes both pixel-level restoration and feature-level regression. Through meticulous juxtaposition with these classical approaches, we are able to holistically appraise the performance and superiority of the proposed method.

Table 2 presents the evaluation results of pre-trained models based on the ViT backbone on three datasets. Sub table (a) shows the evaluation results of fine-tuning experiments, while sub table (b) displays the evaluation results of linear probing experiments.

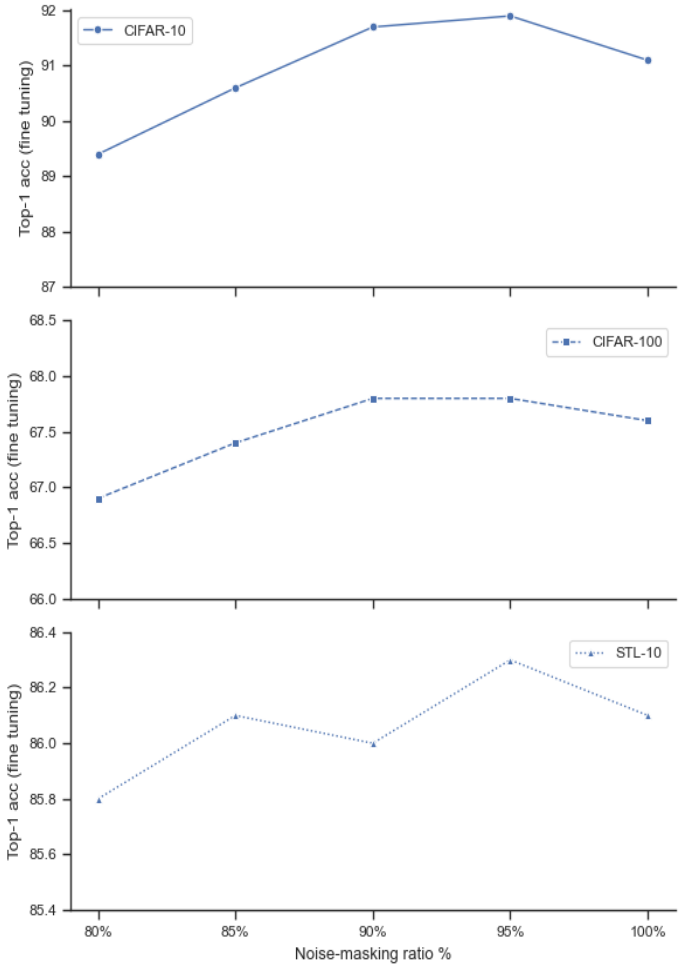

Fig.3. Noise-masking ratio Analysis.

Based on the experimental outcomes, the method we introduced exhibited superior performance within the experimental cohort when full fine-tuning on the CIFAR-10 and STL-10 datasets using the VIT-Small backbone network, with an enhancement of approximately 0.1% in top-1 accuracy. Furthermore, employing the VIT-tiny backbone, our approach manifested an approximate 0.3% elevation in top-1 precision full fine-tuning on the CIFAR-100 dataset. Similarly, utilizing a VIT-based backbone, our methodology demonstrated an approximate 0.1% increment in top-1 performance full fine-tuning on the STL-10 dataset. These observations cogently underscore the efficacy of our approach, achieving results commensurate with those of MAE and Mask Feat. Notably, our method achieved performance parity with MAE pretraining in a mere 500 epochs, as opposed to the 1600 epochs requisite for MAE, signifying a substantial reduction in pretraining duration. In comparison to SdAE, our approach realized a 0.1% gain in top-1 precision, further substantiating the effectiveness of our self-distillation design and image denoising. In comprehensive fine-tuning evaluation experiments across three datasets with diverse ViT architectures, our method consistently outperformed the classical contrastive learning approach, MoCo v3. However, in evaluations employing linear probing, the contrastive learning-based MoCo v3 surpassed MIM-based methodologies. This aligns with prior empirical evidence suggesting the inherent superiority of contrastive learning over MIM-based approaches in the context of linear probing. [13,11].

C. Ablation experiment

Our architecture consists of a teacher encoder, a student encoder, a decoder, and a regression decoder. In our ablation experiments, we investigated the impact of using only the L_y loss, which excludes the teacher network and regression decoder, and using only the L_z loss, which excludes the denoising process performed by the decoder. We compared their linear probing experimental results across three datasets. Table 3 presents the results of our ablation study. When using only the L_y loss, the linear probing experimental results showed a minor decrease of approximately 2%

to 4%, while a significant performance drop was observed when using only the L_z loss. These findings validate the effectiveness of the combination of L_y loss and L_z loss, optimizing the feature extraction capability of the pre-trained model from both pixel-level and high-dimensional feature perspectives.

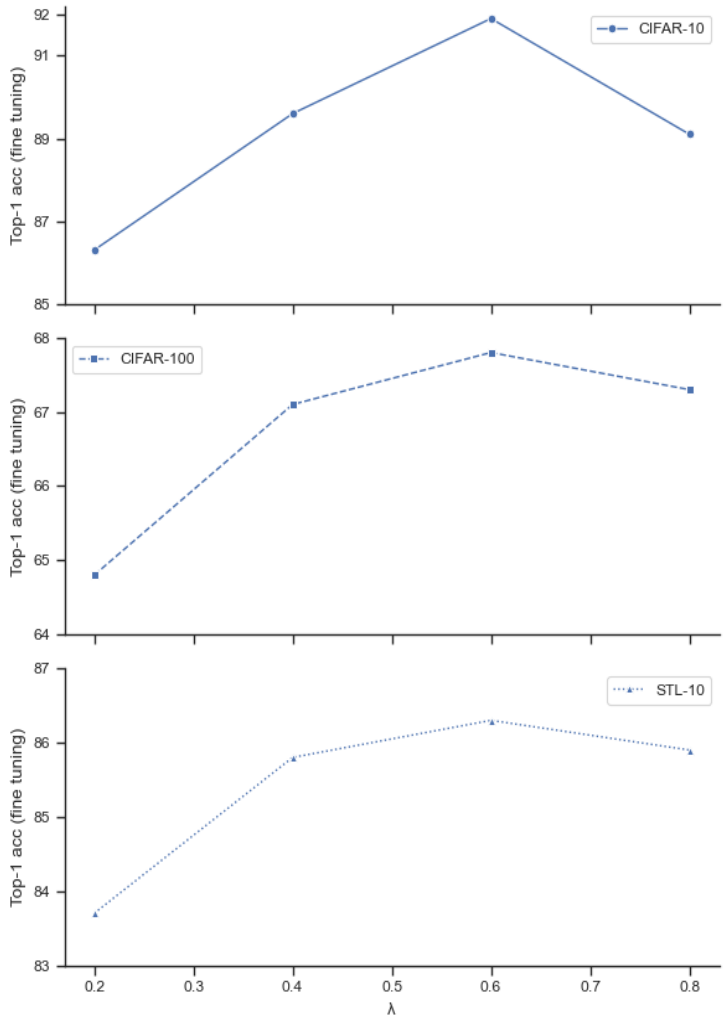

Fig. 4. Optimal λ value.

Table 3. Ablation study on model architecture was conducted, where all models employed the VIT-Small backbone network and underwent 300 epochs of pre-training on CIFAR-10, CIFAR-100, and STL-10 datasets.

Decoder	regression decoder	CIFAR-10	CIFAR-100	STL-10
√	×	74.20	49.74	75.35
×	√	68.93	44.51	61.83
√	√	76.98	53.84	78.56

Noise Masking Ratio: Experimental results in MAE show that the masking rate in masked image modeling can reach up to 75%. Our proposed method, which combines MAE and knowledge distillation, requires a higher noise masking ratio to denoise corrupted images, as depicted in Fig.3. Experimental results obtained from three different datasets indicate that optimal performance is achieved when the noise masking ratio ranges between 90% and 95%. Employing a noise masking ratio that is too low results in overly simplistic proxy tasks for the pre-trained model, leading to a decline in its feature extraction capability.

To investigate the impact of the λ value in Eq. 3. loss function on the feature extraction capability of the pre-trained model, we conducted experiments with various λ values. The fine-tuning results of the pre-trained model are depicted in Fig.4. Experimental results from three datasets indicate that the fine-tuning of the pre-trained model achieves the best results when the λ value is set to 0.6.

5. Conclusion

In this paper, we propose a novel self-supervised learning architecture that combines MAE and feature distillation to optimize the pre-trained model from both pixel-level and high-dimensional feature perspectives. This approach enables the pre-trained model to focus on both global and local information. We conduct experiments on CIFAR-10, CIFAR-100, and STL-10 datasets. The results demonstrate that our method achieves comparable performance to other MIM-based self-supervised learning methods, slightly surpassing the results obtained by MAE. Additionally, our method requires fewer epochs (less than 1600 epochs) to achieve the same top-1 accuracy, resulting in shorter training time compared to MAE. In ablation experiments, we investigate the design framework, validate its rationale, and explore the impact of noise masking ratio on experimental results, finding that our method requires a noise masking ratio between 90% and 95%.

Due to limited computational resources, we validate our method only on relatively small datasets. However, the experimental results thus far are sufficient to demonstrate its effectiveness. In the future, further exploration can be conducted on downstream tasks, such as image recognition and segmentation in the medical field [39,40]. Overall, our method can serve as a complementary approach within the family of self-supervised learning algorithms.

References

[1] Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y. and Schwarzschild, A., 2023. A cookbook of self-supervised learning. arXiv preprint arXiv:2304.12210.

[2] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z. and Wu, Z., 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:2304.01852.

[3] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y. and Dollár, P., 2023. Segment anything. arXiv preprint arXiv:2304.02643.

[4] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A. and Assran, M., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

[5] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020.

[6] Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning." Advances in neural information processing systems 33 (2020): 21271-21284.

[7] He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

[8] Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748.

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[10] Bao, H., Dong, L., Piao, S., & Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

[11] Chen, X., Ding, M., Wang, X., Xin, Y., Mo, S., Wang, Y., ... & Wang, J. (2022). Context autoencoder for self-supervised representation learning. arXiv preprint arXiv:2202.03026.

[12] Dong, X., Bao, J., Zhang, T., Chen, D., Zhang, W., Yuan, L., ... & Guo, B. (2023, June). Peco: Perceptual codebook for bert pre-training of vision transformers. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 552-560).

[13] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 16000-16009).

[14] Wei, C., Fan, H., Xie, S., Wu, C. Y., Yuille, A., & Feichtenhofer, C. (2022). Masked feature prediction for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14668-14678).

[15] R. C. Gonzalez; R. E. Woods, Digital Image Processing, Prentice Hall, Upper Saddle River, NJ., 2002. ISBN 013168728X.

[16] A. Krizhevsky, ‘‘Learning multiple layers of features from tiny images,’’ Tech. Rep., 2009.

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

[18] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., & Sutskever, I. (2020, November). Generative pretraining from pixels. In International conference on machine learning (pp. 1691-1703). PMLR.

[19] Zhou, J., Wei, C., Wang, H., Shen, W., Xie, C., Yuille, A., & Kong, T. (2021). ibot: Image bert pre-training with online tokenizer. arXiv preprint arXiv:2111.07832.

[20] Chen, Xinlei, and Kaiming He. "Exploring simple siamese representation learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 15750-15758. 2021.

[21] Caron, Mathilde, 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, pp. 9650-9660. 2021.

[22] Zbontar, Jure, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. "Barlow twins: Self-supervised learning via redundancy reduction." In International Conference on Machine Learning, pp. 12310-12320. PMLR, 2021.

[23] Assran, Mahmoud, Mathilde Caron, Ishan Misra, Piotr Bojanowski, Florian Bordes, Pascal Vincent, Armand Joulin, Mike Rabbat, and Nicolas Ballas. "Masked siamese networks for label-efficient learning." In European Conference on Computer Vision, pp. 456-473. Cham: Springer Nature Switzerland, 2022.

[24] Jing, Li, Jiachen Zhu, and Yann LeCun. "Masked siamese convnets." arXiv preprint arXiv:2206.07700 (2022).
[25] Chen, Yabo, Yuchen Liu, Dongsheng Jiang, Xiaopeng Zhang, Wenrui Dai, Hongkai Xiong, and Qi Tian. "Sdae: Self-distillated masked autoencoder." In European Conference on Computer Vision, pp. 108-124. Cham: Springer Nature Switzerland, 2022.
[26] Caron, Mathilde, 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, pp. 9650-9660. 2021.
[27] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In ICCV, 2015.
[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In ECCV, 2016.
[29] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In ICLR, 2018.
[30] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In ECCV, 2016.
[31] Zhang, Richard, Phillip Isola, and Alexei A. Efros. "Split-brain autoencoders: Unsupervised learning by cross-channel prediction." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1058-1067. 2017.
[32] Pathak, Deepak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. "Context encoders: Feature learning by inpainting." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2536-2544. 2016.
[33] Xie, Zhenda, 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, pp. 9653-9663. 2022.
[34] Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. and Piot, B., 2020. Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33, pp.21271-21284.
[35] Loshchilov, Ilya, and Frank Hutter. "Sgdr: Stochastic gradient descent with warm restarts." arXiv preprint arXiv:1608.03983 (2016).
[36] Loshchilov, Ilya, and Frank Hutter. "Decoupled weight decay regularization." arXiv preprint arXiv:1711.05101 (2017).
[37] Goyal, Priya, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. "Accurate, large minibatch sgd: Training imagenet in 1 hour." arXiv preprint arXiv:1706.02677 (2017).
[38] Fan, Haoqi, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. "Multiscale vision transformers." In Proceedings of the IEEE/CVF international conference on computer vision, pp. 6824-6835. 2021.
[39] Stirenko, Sergii, Yuriy Kochura, Oleg Alienin, Oleksandr Rokovyi, Yuri Gordienko, Peng Gang, and Wei Zeng. "Chest X-ray analysis of tuberculosis by deep learning with segmentation and augmentation." In 2018 IEEE 38th International Conference on Electronics and Nanotechnology (ELNANO), pp. 422-428. IEEE, 2018.
[40] Xu, J. and Stirenko, S., 2022. Self-supervised Model Based on Masked Autoencoders Advance CT Scans Classification. International Journal of Image, Graphics and Signal Processing, pp.1-9.
[41] J. Xu and S. Stirenko, "Mixup Feature: A Pretext Task Self-Supervised Learning Method for Enhanced Visual Feature Learning," in IEEE Access, vol. 11, pp. 82400-82409, 2023, doi: 10.1109/ACCESS.2023.3301561.

Authors' Profiles

Jiashu Xu received a master's degree from the Department of computing engineering, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute". Now is a Ph.D. student from the same university. His research interests include self-supervised learning, unsupervised learning, computer vision, GAN, and their applications in the medical image domain.

Sergii Stirenko, Head of Computer Engineering Department, Research Supervisor of KPI-Samsung R&D Lab, Head of NVIDIA GPU Education and NVIDIA GPU Research Center, and Professor at National Technical University of Ukraine "Kyiv Polytechnic Institute." Research is mainly focused on artificial intelligence, high-performance computing, cloud computing, distributed computing, parallel computing, eHealth, simulations, and statistical methods. And published more than 60 papers in peer-reviewed international journals.

How to cite this paper: Jiashu Xu, Sergii Stirenko, "Denoising Self-Distillation Masked Autoencoder for Self-Supervised Learning", International Journal of Image, Graphics and Signal Processing(IJIGSP), Vol.15, No.5, pp. 29-38, 2023. DOI:10.5815/ijigsp.2023.05.03