This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Closest Neighbors are Harmful for Lightweight Masked Auto-encoders

[1]Jian Meng, [1]Ahmed Hasssan, [2]Li Yang, [3]Deliang Fan, [4]Jinwoo Shin, and [1]Jae-sun Seo
[1]Cornell University,[2] University of North Carolina at Charlotte, [3] Arizona State University, [4]KAIST
[1]{jm2787, ah2288, js3528}@cornell.edu, [2]lyang50@uncc.edu, [3]dfan12@asu.edu, [4]jinwoos@kaist.ac.kr

Abstract

Learning the visual representation via masked auto-encoder (MAE) training has been proven to be a powerful technique. Transferring the pre-trained vision transformer (ViT) to downstream tasks leads to superior performance compared to conventional task-by-task supervised learning. Recent research works on MAE focus on large-sized vision transformers (>50 million parameters) with outstanding performance. However, improving the generality of the under-parametrized lightweight model has been widely ignored. In practice, downstream applications are commonly intended for resource-constrained platforms, where large-scale ViT cannot easily meet the resource budget. Current lightweight MAE training heavily relies on knowledge distillation with a pre-trained teacher, whereas the root cause behind the poor performance remains under-explored. Motivated by that, this paper first introduces the concept of "closest neighbor patch" to characterize the local semantics among the input tokens. Our discovery shows that the lightweight model failed to distinguish different local information, leading to aliased understanding and poor accuracy. Motivated by this finding, we propose NoR-MAE, a novel MAE training algorithm for lightweight vision transformers. NoR-MAE elegantly repels the semantic aliasing between patches and their closest neighboring patch (semantic centroid) with negligible training cost overhead. With the ViT-Tiny model, NoR-MAE achieves up to 7.22%/3.64% accuracy improvements on ImageNet-100/ImageNet-1K datasets, as well as up to 5.13% accuracy improvements in tested downstream tasks. https://github.com/SeoLabCornell/NoR-MAE

1. Introduction

Starting from the early exploration with contrastive learning [6, 18], learning powerful and generic vision or semantic representation has been the focal point of all prior self-supervised learning (SSL) algorithms across various machine learning domains [2, 8, 28, 30] and vision-language model [15, 26]. The superiority of SSL is demonstrated by the capability of extracting transferable knowledge from unlabeled datasets [26, 33]. Recent efforts on masked autoencoder (MAE) [19] train the vision transformer [11] by encouraging the model to understand the semantic knowledge of the masked patches throughout the reconstruction-based self-supervised learning. Particularly, the evolution from contrastive learning [16, 18, 36] to MAE-based training [17, 19, 34, 37] shows consistent dominance by large-scale models to ensure on-par or even better accuracy compared to supervised learning, along with superior performance in various downstream applications. In practice, downstream small-scale tasks are commonly deployed on edge devices with limited resources. Therefore, large-sized vision transformers (ViTs) pre-trained by MAE [19] are **sub-optimal candidates** for resource-constrained downstream deployment and hardware platforms. Unfortunately, recent state-of-the-art (SoTA) methods that train lightweight ViTs with MAE from scratch lead to poor performance and insufficient transferability. For example, the MAE-trained lightweight ViT-Tiny model only achieves 66.6% top-1 accuracy (with 200-epoch pre-training [3, 19, 37]) on the

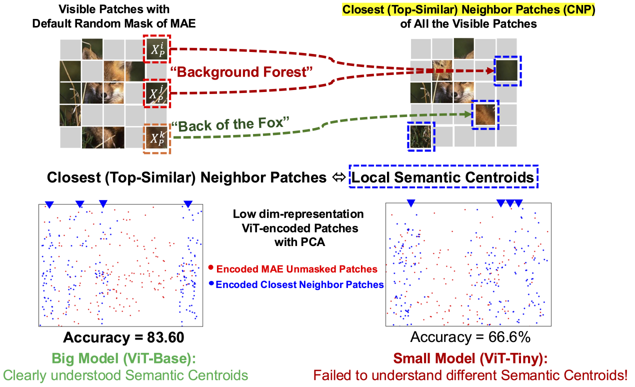
Visible Patches with
Default Random Mask of MAE
Closest (Top-Similar) Neighbor Patches (CNP)
of All the Visible Patches
X_P^i
X_P^j
X_P^k
"Background Forest"
"Back of the Fox"
Closest (Top-Similar) Neighbor Patches ⇔ Local Semantic Centroids
Low dim-representation
ViT-encoded Patches
with PCA
● Encoded MAE Unmasked Patches
● Encoded Closest Neighbor Patches
Accuracy = 83.60
Big Model (ViT-Base):
Clearly understood Semantic Centroids
Accuracy = 66.6%
Small Model (ViT-Tiny):
Failed to understand different Semantic Centroids!

Figure 1. **Top:** **C**losest **N**eighbor **P**atches (CNP) characterizes the "Local Semantic Centroid" among the input tokens. **Bottom:** Low-dimensional projection (via PCA) of the transformer-encoded patches. Unlike the large-sized transformer model (e.g., ViT-Base), lightweight models (e.g., ViT-Tiny) cannot separate the "semantic difference" between different CNPs.

ImageNet-1K dataset [3], which is ∼10% lower than the supervised learning counterpart. Evidently, the superiority of MAE [19] with the large model failed to be maintained in the lightweight ViTs [11]. Given the diverse downstream tasks and urgent need for energy-efficient foundation models, unleashing the power of MAE and enhancing the versatility of lightweight models is desired.

Prior works enhance the performance of the lightweight transformer models by heavily relying on knowledge distillation (KD) [20] for both supervised [29] and self-supervised learning paradigms [3, 21, 38]. For instance, the recent DMAE work [3] combines MAE [19] with KD and transfers the pre-trained knowledge from a ViT-Base teacher down to the lightweight ViT model. DMAE [3] chooses the 3/4 depth of both student and teacher ViT models as the bonding portal for feature alignment and knowledge distillation. The heuristic design of DMAE cannot be easily extended to other encoder models. Furthermore, distilling the knowledge on the fly introduces an additional projector [3] or logits mask [31], which further elevates the complexity and computation cost of the entire training process. In addition to DMAE, other recent works further increase the complexity of the KD-based MAE by employing multi-layer distillation [38] or complex generic-to-specific distillation [21]. All of these prior works [3, 21, 31, 38] rely on the **brute force** approach to enhance the model performance with increasingly complex training schemes, which results in inflated GPU memory and training time.

However, none of these prior works can collectively achieve 1) the capability of training lightweight model **from scratch** without a large-sized teacher model, 2) negligible training cost overhead compared to the vanilla MAE, and 3) clear insights that cause the poor performance of lightweight MAE. To overcome this, the key questions to address are:

① What is the bottleneck of MAE with under parametrized small vision transformer models?

② How to efficiently train a high performance lightweight model from scratch via MAE?

To answer ①, our investigation shows that the lightweight model failed to understand different local semantic information among the input tokens and lead to poor performance. We characterize different local semantics by introducing the concept of **closest neighborhood patch** (CNP) (Figure 1), which carries the most similar semantics of a given input patch. Since multiple patches can share one CNP (e.g., Background Forrest), different CNPs represent various local semantics across the entire input sequence. Unlike large-sized models, the under-parameterized model falls into the semantic aliasing between different CNPs with poor accuracy, as indicated in Figure 1 and Section 3.3.

Motivated by the findings in ①, we propose the **N**eighb**o**r **R**epelling **M**asked **A**utoencoder (NoR-MAE) to resolve ②, a novel self-supervised learning algorithm designed for training powerful lightweight vision transformers. NoR-MAE trains the powerful small vision transformer by penalizing the similarity between the unmasked patches and CNPs. Different from prior works, NoR-MAE directly trains the vision transformer from scratch without introducing the large-sized pre-trained teacher or heuristically designed distillation scheme. The proposed method achieves up to **7.22%** and **3.64%** accuracy improvements on ImageNet-100 and ImageNet-1K datasets, with negligible training time and memory overhead compared to the vanilla MAE. The major contributions of our work are:

- **Simplicity:** NoR-MAE does not require teacher pretraining or heuristic knowledge distillation to train a powerful lightweight ViT.
- **High Performance:** Compared to the baseline MAE, NoR-MAE achieves **7.22%** and **3.64%** accuracy improvements on ImageNet-100 and ImageNet-1K datasets with the ViT-Tiny model, achieving the new SoTA performance on the lightweight MAE training.
- **Transferrability:** Different from prior works, which mainly focus on fine-tuning the performance on the pretraining dataset (e.g., ImageNet), NoR-MAE is also evaluated with the downstream fine-tuning and linear evaluation benchmarks, with significantly improved performance compared to prior SSL baselines and the vanilla MAE.
- **Rationality:** Besides the superior performance, **for the first time**, NoR-MAE reveals the semantic aliasing issue caused by the neighboring patches, which is proved to be the reason that hinders efficient training and representation learning of lightweight ViT models.

2. Related Work

Empowering a deep neural network model with strong and transferable knowledge has been widely investigated, especially in self-supervised learning. Early research works in contrastive learning (CL) introduce the learning paradigm with “positive” and “negative” sample pairs [6, 18], together with the alignment-repellent strategy embedded into InfoNCE [7, 18, 25] and NT-Xent loss [6] as the learning objective. The early success of CL motivated the successors to investigate the potential of different training and sampling strategies, inducing the standpoint of asymmetrical learning [16, 24], knowledge distillation [13], and the latent-dimension correlation [36].

While the core idea of CL is training the model to understand the similarities and differences between different augmentations, the emergence of masked autoencoder (MAE) [19] training reconsiders SSL from a different perspective, which is perfectly suitable for vision transformers. Encouraging the encoder-decoder to reconstruct the randomly sparsified patches allows the model to understand the semantics. Due to the powerful visual representation, MAE-based pretaining exhibits strong performance on

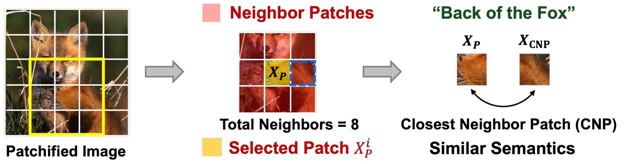

Figure 2. Closest Neighbor Patches (CNP) within the image.

both backbone datasets (e.g., ImageNet-1K) and downstream transfer learning tasks. Subsequent research works further improved the performance of MAE from the perspectives of contrastive learning [27, 37], cross-attention [14, 17], mixed feature learning [23], or intricate tokenizer [12].

Across all the different MAE training methods, elevating the performance of the large-sized models (e.g., ViT-Large) has been the primary focus of exploration, while improving the poor performance of the small-sized model (e.g., ViT-Tiny) has been largely ignored. Given that the downstream tasks are well-suited for resource-constrained edge devices, the imbalanced research focus of the prior works largely hinders the efficiency and transferability of MAE [19]. The practical needs and the imperfection of lightweight MAE motivate the recent research works to investigate MAE training strategies with small models. Following the protocol from supervised distillation [20], knowledge distillation-based MAE [3, 21] is proposed to enhance the performance of the lightweight student. However, the learning scheme of the single-layer or multi-layer distillation is established based on the heuristically designed distillation between intermediate features. Compared to the supervised knowledge distillation, recent studies largely complicate the overall cost of training lightweight ViT models with either heuristic design [3], multi-phased distillation [21], or distillation with extensive fine-tuning [31]. In particular, employing the pre-trained large ViT model as the teacher introduces additional GPU memory usage and time cost toward the total training efforts. Furthermore, training the interconnect projectors between the student and teacher exacerbates the training complexity even further. More importantly, the knowledge distillation-based MAE inherits the essence of KD from supervised learning, but the insights of training lightweight ViT **from scratch** via MAE remain largely unexplored.

3. Proposed Method

We propose the Neighbor-Repelling MAE (NoR-MAE). In this section, we will 1) introduce the concept of the Closest Neighbor Patch into MAE training, 2) present the proposed NoR-MAE training, and 3) unravel the rationality of NoR-MAE by revealing the insights that cause the poor performance of the directly-trained lightweight ViT model.

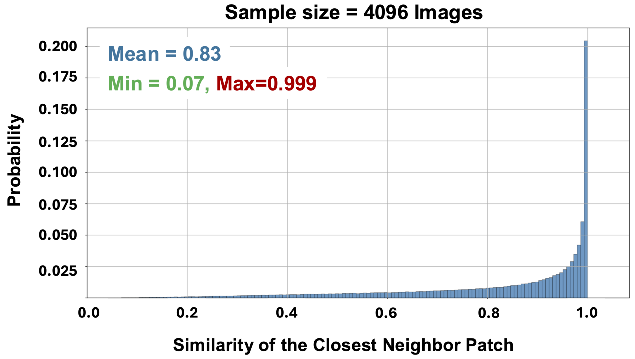

Figure 3. Cosine Similarity between each patch and their Closest Neighbor Patch across a subset of ImageNet-1K with 4096 images.

3.1. Closest Neighbor Patch (CNP)

Let $X \in \mathbf{R}^{H \times W \times C}$ be an input image to the model, where H, W, and C represent the height, width, and channel of the image. Vision transformer (ViT) reformulates the input by decomposing the image X into a sequence of non-overlapped patches with the pre-defined patch size $P \times P$ and the total number of patches (tokens) K. All the patches are further encoded into the embeddings with dimension D, which is the input of the subsequent transformer blocks.

For a given input patch X_P, there are naturally 8 neighbors that share a common edge with X_P, as shown in Figure 2(a). We define the **Closest Neighbor Patch** (CNP) X_{CNP} by selecting the neighbor that has the Top-1 cosine similarity with X_P. In other words, X_P and X_{CNP} are sharing the highest degree of semantically meaningful relations.

Shared Closest Neighbor Patch. In the example of Figure 2, two patches X_P^i and X_P^j partially share the neighborhoods in between, while the CNP of X_P^i and X_P^j happen to be identical as X_{CNP}^{ij}. Naturally, X_{CNP}^{ij} carries the shared semantic information ("Background Forest") for **both** X_P^i and X_P^j, but the degree of similarity between $(X_P^{\mathbf{i}}, X_{CNP}^{ij})$ and $(X_P^{\mathbf{j}}, X_{CNP}^{ij})$ are different.

Closest Neighbor Patch $\neq$ Identical Patch. Although X_P and X_{CNP} are sharing the highest degree of similarity, the semantics between X_P and X_{CNP} are **not** identical. As shown in Figure 3, the cosine similarity between all the patches and their closest neighbor in ImageNet-1K images varies from 0.07 to 0.999, with an average of 0.83. In other words, the highly similar neighboring patches also contain the implicit semantical difference.

Difficulty of Lightweight Models to Understand the Different Semantics Carried by CNP. With MAE, the input patches of the transformer models are randomly masked with a pre-defined sparsity ratio τ (e.g., 75%), as illustrated in Figure 4, left. Naturally, each unmasked visible patch has a **Closest Neighbor Patch (CNP)**. As a result, the CNP

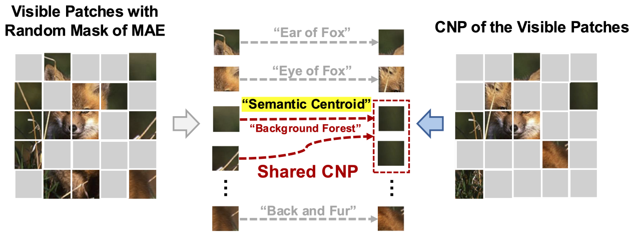

Figure 4. The shared Closest Neighbor Patch (CNP) characterizes the estimated local semantic centroids of the input sequence.

of the unmasked patches formulates another sparse input image (Figure 4, right). From the perspective of the entire image, visible patches and their CNPs are close neighbors, while different visible patches may share the same closest neighbor, which estimates the local and regional semantics of the input sequence. However, as described in Figure 1 and Section 1, understanding the semantic difference between different CNPs is the bottleneck for the under-parameterized models. Furthermore, our extended experiments in Figure 7 show that the **key difference** between the big (e.g., ViT-Base) and small (e.g., ViT-Tiny) model is understanding and separating the semantic differences between CNPs. Therefore, the challenge of lightweight MAE is clear:

In MAE training, how to facilitate the lightweight model to understand different local semantics?

3.2. Proposed Method: NoR-MAE

We propose the Neighbor-Repelling training for self-supervised MAE learning (NoR-MAE). The overall objective is to train a powerful, lightweight vision transformer by encouraging the model to avoid the aliased relation between different local semantics of the input. We follow the standard protocol of the vanilla MAE with the encoder f_θ and decoder g_ϕ, where both f_θ and g_ϕ are vision transformers with the same embedding dimension D.

Given the masking ratio τ, we first generate the masked input sequence x^τ and x^τ_{CNP} that are embedded from the unmasked patches with the corresponding closest neighbor patches, respectively. During the forward pass of each iteration, we parallely encode and decode the masked sequence x^τ and CNP sequence x^τ_{CNP} by f_θ and g_ϕ, generating the reconstructed embedding Z and Z_{CNP}, respectively. In particular, f_θ and g_ϕ are served as the siamese encoder and decoder (with stop gradient) during the forward pass of Z_{CNP}.

Overall, each forward pass will generate two pairs of decoded tokens resulting from x^τ and x^τ_{CNP}.

Following the standard training protocol of MAE, we first compute the $\mathcal{L}_2$ reconstruction loss between the decoded sequence Z and the ground truth. Subsequently, we compute the matrix multiplication between the normalized Z and Z_{CNP}. As shown in Figure 5, the encoder f_θ and decoder g_ϕ are shared for both branches, while the stop gradient is enabled for the CNP input. Given the size of $[N, K, D]$, the matrix multiplication between $\bar{Z}$ and transposed $\bar{Z}_{\text{CNP}}$ is performed along the token dimension K. The resultant matrix is further scaled down by K and takes the average along the batch dimension N. Mathematically, we have:

$$C_{\text{neighbor}} = \frac{1}{N}\sum_{N}\left(\frac{1}{K}(\bar{Z})^{T}\cdot\bar{Z}_{\text{CNP}}\right) \tag{1}$$

Where C_{neighbor} represents the resultant relation matrix with the size of $D \times D$. Each entry $c_{i,j}$ of C_{neighbor} characterizes the correlation between the randomly-unmasked input sequence (output logits Z) and the corresponding CNP sequence (output logits Z_{CNP}) between embedding dimension i and j. To minimize the semantic aliasing between the unmasked tokens and their CNPs, we decorate different embeddings by penalizing the off-diagonal terms of C_{neighbor}:

$$\mathcal{L}_{\text{NoR}} = \sum_{i}\sum_{i\neq j} C^{i,j}_{\text{neighbor}} \tag{2}$$

Essentially, the proposed NoR-Loss ($\mathcal{L}_{\text{NoR}}$) minimizes semantic aliasing by preventing the model encoding the visible patches and CNP as aliased embeddings. It has been shown in Figure 3 that patches and their CNPs share similar but non-identical semantics. We balance such a tradeoff of "similarity-discrepancy" by decorrelating the off-diagonal embeddings only. The overall loss is the combination of the standard $\mathcal{L}_2$ reconstruction loss and the proposed Neighbor-Repelling Loss $\mathcal{L}_{\text{NoR}}$ scaled by a hyperparameter λ:

$$\mathcal{L}_{\text{total}} = ||\text{Ground Truth} - \bar{Z}||_2 + \lambda \cdot \mathcal{L}_{\text{NoR}}(\bar{Z}, \bar{Z}_{\text{CNP}}) \tag{3}$$

More importantly, decorrelating embeddings eventually separate the aliased semantics between different CNPs and their corresponding "local semantic centroids" (Section 3.3). In other words, the **embedding-level anti-aliasing** regularization of the proposed NoR-Loss facilitates the understanding on the **representation (token) level**, leading to the largely improved model performance.

Decorrelating Embeddings vs. Decorrelating Tokens. Although the proposed NoR-MAE selects the Top-1 neighbor (the closest neighbor patch (CNP)) as the candidate for decorrelation, **the CNP patches can be overlapped or shared between each other.** For instance, the closest neighbor patch of patch X_i could also be the Top-2/3/4 … 8 neighbor of patch X_j. Minimizing the semantic aliasing by naïvely decorrelating tokens leads to collapsed performance and unsuccessful training, as shown in Figure 6.

NoR-MAE vs. Contrastive Learning. Empowering representation learning via invariance-covariance optimization has been investigated in Barlow Twins [36]. Recently proposed U-MAE [37] combines the contrastive learning and MAE [19] by introducing the **random** asymmetry masks

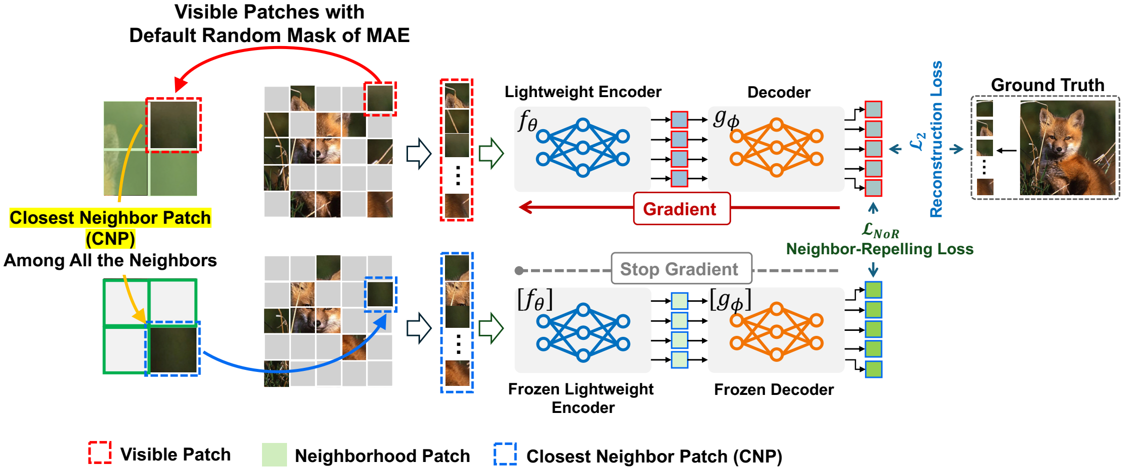

Figure 5. Overview of the proposed NoR-MAE algorithm.

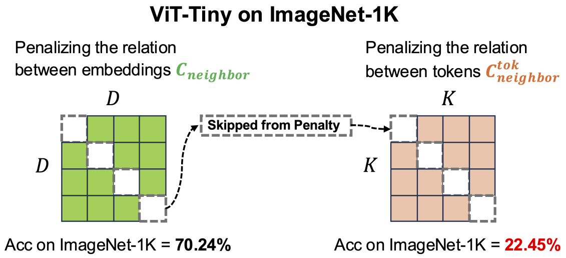
ViT-Tiny on ImageNet-1K
Penalizing the relation
between embeddings $C_{neighbor}$
Penalizing the relation
between tokens $C^{tok}_{neighbor}$
D
D
K
K
Skipped from Penalty
Acc on ImageNet-1K = 70.24%
Acc on ImageNet-1K = 22.45%

Figure 6. Different relation penalty of the NoR-Loss (λ=5e-5).

during the forward pass of the training. On the contrary, the proposed NoR-MAE exploits the semantic aliasing issue by constructing the asymmetry between the visible patches and the corresponding closest neighbors, instead of relying on the randomness. As a result, NoR-MAE outperforms U-MAE [37] in various sizes of models (Section 4).

In the context of contrastive learning, the similar samples (patches) are considered as "positive samples" which are intended to be aligned by the algorithms [35]. However, the important discovery of NoR-MAE points out that semantic aliasing among the closest neighbors is **harmful** for the lightweight masked autoencoder training, which is not properly addressed by prior works. Moreover, unlike Barlow Twins [36], NoR-MAE does not need the embedding alignment along the diagonal terms due to the well-established similarity between the visible patches and their CNPs.

NoR-MAE Does Not Need a Teacher Model. We would like to highlight that the proposed NoR-MAE is entirely different from knowledge distillation-based MAE. NoR-MAE does not require a pre-trained teacher [3] or multi-stage distillation [21] during training. The lightweight encoder is trained from scratch following the standard MAE protocol [19]. Although the CNPs and the original unmasked patches are separately generated, the encoding-decoding process remains the same as the vanilla MAE. For example, given the 75% mask ratio with 25% visible patches, the equivalent input sparsity of NoR-MAE is 50%. The overhead of the NoR-MAE is minimal compared to the vanilla MAE, as profiled in the supplementary.

3.3. Rationality of NoR-MAE

As shown in Figure 2 and Figure 4, the proposed concept of Closest Neighbor Patch (CNP) characterizes the **local semantics** of the image. For instance, the patches that contain "Background Forest" or "Back of the Fox" are represented by separate CNP patches. However, the lightweight model failed to learn the semantic difference between CNPs.

We prove the significance of the CNP semantics by analyzing the original unmasked features and CNP features encoded by the trained encoder f_θ. Specifically, we first normalize the encoded features and then compress the feature dimension down to 2-D via the Principal Component Analysis (PCA) [1]. As shown in Figure 7 and Figure 8, the lightweight models (ViT-Tiny and ViT-Small) trained by the vanilla MAE [19] **failed** to distinguish the semantical difference between different closest neighbor patches (CNP, blue dots), regardless of the patch sizes. Different semantics of CNP are heavily mixed together in the low dimensional space. Moreover, CNP is critical for MAE training regardless of the input patch sizes and the corresponding sequence length. The distorted and aliased understanding of the local semantics leads to the sub-optimal performance of MAE, as shown in Table 1 with the empirical verification on ImageNet-100 (224×224) with ViT-Tiny.

On the contrary, the model trained by NoR-MAE successfully understands the semantic difference between different CNP features, leading to largely improved accuracy on the lightweight ViT models across different patch sizes. In the meantime, the large-sized model (e.g., ViT-Base) exhibits

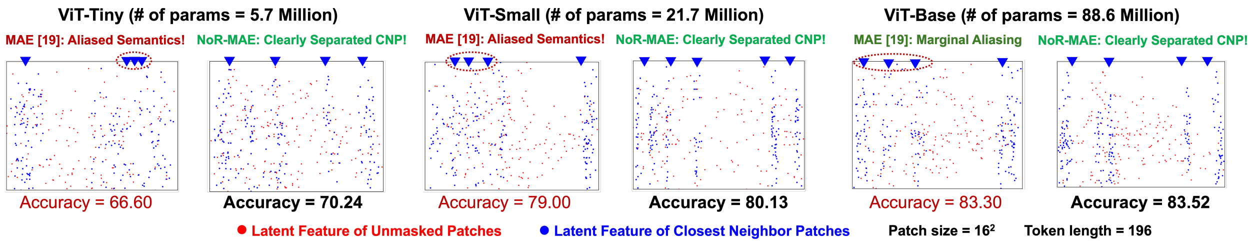

Figure 7. Low dimensional features of the unmasked patches (red dots) and the closest neighbor patches (CNP) (blue dots) with the **same input** from ImageNet-1K dataset. NoR-MAE clearly separates the semantic information among different CNP, with improved accuracy.

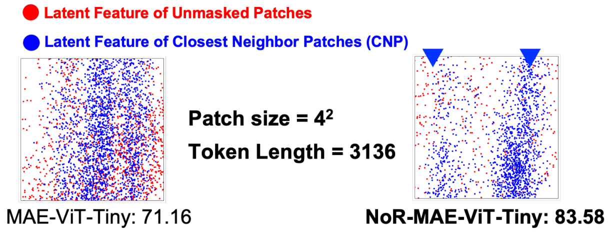

Figure 8. Low dimensional features of the unmasked patches (red dots) and the CNP (blue dots) encoded by ViT-Tiny with patch size = 4×4 (sequence length = 3136), evaluated on ImageNet-100.

Table 1. NoR-MAE consistently outperforms the MAE [19] across different input patch sizes (evaluated on the ImageNet-100 dataset).

Masking Ratio	Input Patch Size	MAE [19]	NoR-MAE (This work)
75%	2^2	71.08%	**83.56% (+12.48%)**
75%	4^2	71.16%	**83.58% (+12.42%)**
75%	8^2	71.27%	**79.24% (+7.97%)**
75%	16^2	71.04%	**78.26% (+7.22%)**
75%	32^2	68.42%	**74.73% (+6.31%)**

a better understanding and stronger separation of different semantics with the vanilla MAE [19].

In summary, the proposed NoR-MAE successfully mitigates the performance gap between model sizes, facilitating the learnability of the transformer model with a stronger understanding of the semantical relations from the input, leading to improved accuracy with **both** lightweight and large-sized transformer models.

4. Experimental Results

4.1. ImageNet Classification

For image classification tasks, we follow the standard protocol of MAE [19] by training the lightweight encoder on the ImageNet-1K and ImageNet-100 datasets with the batch size of 4,096 and 2,048, and subsequently performing the standard fine-tuning [19]. Specifically, the model is pre-trained by **200 and 400 epochs** with the initial learning rate set to 1.5e-4, together with the weight decay value of 0.05. The penalty coefficient λ of the proposed NoR Loss is set to 5e-5, and the optimality of the selected penalty level is demonstrated in Table 5. The mask ratio is fixed at 75% based on the findings in the vanilla MAE [19]. We implement the standard fine-tuning protocol with 100 epochs as the vanilla MAE [19]. The detailed experimental setup (pre-train and downstream) are summarized in the supplementary, together with the report of GPU memory usage and training cost.

Table 2 and Table 3 summarize the performance of the NoR-MAE on ImageNet-100 and ImageNet-1K datasets. Compared to the vanilla MAE [19], NoR-MAE largely improves the performance on top of the MAE baseline [19] with up to 7.22% and 3.64% accuracy on ImageNet-100 and ImageNet-1K datasets, respectively. For the large-sized vision transformer training (e.g., ViT-Base), NoR-MAE still maintains the superiority compared to the recent high-performance MAE training (Table 3).

More importantly, NoR-MAE improves the model performance **marginal** training cost overhead. NoR-MAE trains the model with the standard MAE protocol without introducing 1) a pre-trained teacher [3] or 2) other additional architectures [27] or clustering modules [12]. As profiled in the supplementary, the performance boost reward by the NoR-MAE only exhibits up to **8%** memory overhead and **2%** additional training time. Compared to the distillation-based MAE [3], NoR-MAE outperforms DMAE [3] with ∼1% accuracy improvement and **50%** less training cost.

We further validate the NoR-MAE on the larger ViT-Base model. The proposed algorithm consistently achieves better performance compared to the MAE baseline and the recent ImageNet-100 SoTA method without introducing any additional clustering efforts or special tokenizer [12].

Finally, NoR-MAE is also suitable for the MAE with supervision (SupMAE) [22]. Introducing supervised learning into MAE does not impact the effectiveness of the proposed NoR-MAE, while the NoR-MAE further enhances the performance of SupMAE [22], as shown in Table 2.

Comparison with the Distillation-based MAE. We would like to **highlight** that the proposed NoR-MAE can achieve on-par or even better performance compared to the

Table 2. Fine-tuning and semi-supervised learning accuracy (%) of the ViT encoder pre-trained by MAE [19] and recent SoTA lightweight MAE on ImageNet-1K dataset (with 200 epochs and 400 epochs pre-training).

ViT Model	# of Params	Method	Teacher Model	Pre-training Epoch	Supervised Fine-tuning Accuracy (%)
ViT-Tiny	5.7M	MAE [19]	-	200	66.60
		MAE [19]	-	400	67.63
		Sup MAE [22]	-	200	67.88
		Sup MAE [22]	-	400	68.91
		MoCo-V3 [18]	-	200	70.09
		DINO [5]	-	300	68.17
		SimMIM [34]	-	200	66.08
		DMAE [3]	ViT-B	200	70.00
		Sup NoR-MAE (This work)	-	200	**70.63 (+2.75)**
		NoR-MAE (This work)	-	200	**70.24 (+3.64)**
		NoR-MAE (This work)	-	400	**71.18 (+3.55)**
ViT-Small	21.7M	MAE [19]	-	200	79.00
		MAE [19]	-	400	80.11
		Sup MAE [22]	-	200	79.27
		CrossMAE [14]	-	400	79.30
		MoCo-V3 [18]	-	200	79.46
		DMAE [3]	ViT-B	200	79.30
		Sup NoR-MAE (This work)	-	200	**80.16 (+0.89)**
		NoR-MAE (This work)	-	200	**80.13 (+1.03)**
		NoR-MAE (This work)	-	400	**81.04 (+0.93)**
ViT-Base	88.6M	MAE [19]	-	200	83.30
		Sup MAE [22]	-	200	83.60
		DMAE [3]	-	200	84.00
		CrossMAE [14]	-	200	83.60
		MoCo-V3 [3]	-	200	83.14
		DINO [5]	-	200	82.80
		SimMIM [34]	-	200	83.50
		U-MAE [37]	-	200	83.00
		NoR-MAE (This work)	-	200	**83.42 (+0.12)**
		Sup NoR-MAE (This work)	-	200	**83.70 (+0.10)**
		NoR-DMAE (This work)	-	200	**84.22 (+0.22)**

distillation-based MAE, **while significantly reducing the training resources**, in terms of GPU training time and memory usage. As shown in Table 2, NoR-MAE directly trains the lightweight model from scratch, achieving on-par or even better performance compared to the recent distillation-based MAE training, which yet consumes massive GPU resources.

4.2. Transfer Learning on the Downstream Tasks

We fine-tune the pre-trained ViT-Tiny and ViT-Small models (from Table 2) on the downstream vision tasks with **100 epochs.** As shown in Table 4, the proposed NoR-MAE algorithm outperforms the MAE baseline [19] among all the downstream tasks with up to 5.13% accuracy improvements. The outstanding downstream performance indicates the capability of learning strong representation through NoR-MAE.

4.3. Ablation Study

Intensity of Neighbor-Repelling (λ). We evaluate the impact of the penalty level λ with different values. As shown in Table 5, the proposed NoR-MAE algorithm achieves the best performance with λ = 5e-5. In other words, the repelling between the embedding features of the visible patches and the CNP should be properly controlled. Over-penalized semantic relation leads to sub-optimal performance.

Impact of Masking Ratio. In addition to the 75% masking ratio that has been used in all prior works as the default setting, we evaluate the impact of different input masking ratios, varying from 50% to 85% sparsity. Among different input masking ratios, the proposed NoR-MAE consistently outperforms the vanilla MAE [19] with lightweight vision

Table 3. Fine-tuning accuracy (%) of the vision transformer encoder pre-trained by MAE and recent SoTA MAE methods on ImageNet-100.

ViT Model	# of Parameters	Pre-training Method	Pre-training Epochs	Supervised Fine-tuning Accuracy (%)
ViT-Tiny	5.7M	MAE [19]	200	71.04
		DiNO-Tiny [5, 9]	200	63.04
		CrossMAE [14]	200	70.82
		NoR-MAE (This work)	200	**78.26 (+7.22)**
ViT-Small	21.7M	MAE [19]	200	81.82
		MaskFeat HOG [32]	200	82.80
		PeCo [10]	200	83.60
		NoR-MAE (This work)	200	**84.28 (+2.46)**
ViT-Base	88.6M	MAE [19]	200	86.80
		CrossMAE [14]	200	86.29
		U-MAE [37]	200	86.80
		BEiT [4]	200	86.10
		NoR-MAE (This work)	200	**87.56 (+0.76)**

Table 4. Fine-tuning accuracy (%) of the NoR-MAE on the downstream tasks with the lightweight ViT pre-trained on the ImageNet-1K.

ViT Model	Method	CIFAR-10	CIFAR-100	Flowers	Pets	DTD	Food	Caltech-101	Aricraft	ADE20K (mIoU)
ViT-Tiny	MAE [19]	95.01	78.74	85.80	82.47	55.07	75.63	68.05	64.60	26.54
	NoR-MAE (This work)	**95.98**	**80.13**	**88.71**	**85.29**	**60.18**	**78.71**	**73.18**	**67.07**	**33.54**
ViT-Small	MAE [19]	97.50	84.83	91.49	91.14	62.31	81.57	80.85	55.90	41.14
	NoR-MAE (This work)	**98.17**	**85.57**	**91.85**	**91.70**	**64.79**	**83.15**	**82.08**	**57.41**	**42.99**

Table 5. Impact of different penalty levels (λ) on Neighbor Repelling on the ImageNet-100 dataset.

Penalty Level (λ)	0.0 (Baseline)	1e-5	5e-5	1e-4	5e-4
ViT-Tiny	70.24	76.20	**78.26**	77.89	72.17
ViT-Small	81.82	82.14	**84.28**	**84.21**	81.83

Table 6. Impact of different input masking ratios. NoR-MAE shows consistent performance improvements on the ImageNet-1K dataset with ViT-Tiny model.

Patch-wise Masking Ratio	Baseline MAE [19]	**NoR-MAE (This work)**
50%	65.43%	**69.27%**
75%	66.60%	**70.24%**
85%	66.37%	**70.16%**

transformers, as shown in Table 6. In other words, semantic aliasing is persistent in lightweight transformer models regardless of the information density in the input sequence.

NoR-MAE with extended training effort We further report the performance of the NoR-MAE algorithm with 1,000 and 1,600 epoch training on the ImageNet-100. As shown in Table 7, the extended training effort can further boost up the performance of the NoR-MAE-trained ViT-Tiny (5.2M) model to 83.96%, while consistently outperforming MAE.

Table 7. Performance of NoR-MAE with extended training effort.

Training Epochs	MAE on ImageNet-100 (%)	**NoR-MAE on ImageNet-100 (%)**
200	71.04	**78.26**
1000	73.22	**82.97**
1600	74.29	**83.96**

5. Conclusion

In this paper, we propose NoR-MAE, a novel self-supervised learning algorithm designed for lightweight vision transformers via masked autoencoder training. We first introduce the concept of closest neighbor patches (CNP) into the MAE training, which is a critical concept in lightweight SSL. On top of that, we propose Neighbor-Repelling Loss for MAE training. The proposed method trains the lightweight model from scratch without using a pre-trained teacher model. NoR-MAE achieves the new state-of-the-art performance on direct lightweight MAE training with largely improved performance on both pre-training and downstream vision tasks. Finally, our discovery of the aliased semantic relations provides valuable insights regarding lightweight masked autoencoder learning.

6. Acknowledgment

This work is supported in part by the National Science Foundation under grant 2403723, and CoCoSys Center in JUMP 2.0, an SRC program sponsored by DARPA.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 5

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning (ICML)*, 2022. 1

[3] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 5, 6, 7

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations (ICLR)*, 2022. 8

[5] 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 International Conference on Computer Vision (ICCV)*, 2021. 7, 8

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020. 1, 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved Baselines with Momentum Contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021. 1

[9] Victor Guilherme Turrisi da Costa, Enrico Fini, Moin Nabi, Nicu Sebe, and Elisa Ricci. solo-learn: A library of self-supervised methods for visual representation learning. *Journal of Machine Learning Research*, 2022. 8

[10] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu, and Baining Guo. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2023. 8

[11] 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, 2

[12] Tianqi Du, Yifei Wang, and Yisen Wang. On the role of discrete tokenization in visual representation learning. In *International Conference on Learning Representations (ICLR)*, 2023. 3, 6

[13] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[14] Letian Fu, Long Lian, Renhao Wang, Baifeng Shi, Xudong Wang, Adam Yala, Trevor Darrell, Alexei A Efros, and Ken Goldberg. Rethinking patch dependence for masked autoencoders. *arXiv preprint arXiv:2401.14391*, 2024. 3, 7, 8

[15] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024. 1

[16] 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 (NeurIPS)*, 2020. 1, 2

[17] Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 3

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 7

[19] 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 (CVPR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3

[21] Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3, 5

[22] Feng Liang, Yangguang Li, and Diana Marculescu. Supmae: Supervised masked autoencoders are efficient vision learners. *arXiv preprint arXiv:2205.14540*, 2022. 6, 7

[23] Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, and Hongsheng Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[24] Jian Meng, Li Yang, Kyungmin Lee, Jinwoo Shin, Deliang Fan, and Jae sun Seo. Slimmed asymmetrical contrastive learning and cross distillation for lightweight model training. In *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 2

[25] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, 2021. 1

[27] Bowen Shi, Xiaopeng Zhang, Yaoming Wang, Jin Li, Wenrui Dai, Junni Zou, Hongkai Xiong, and Qi Tian. Hybrid distillation: Connecting masked autoencoders with contrastive learners. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. 3, 6

[28] Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[29] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning (ICML)*, 2021. 2

[30] Jue Wang, Gedas Bertasius, Du Tran, and Lorenzo Torresani. Long-short temporal contrastive learning of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (NeurIPS)*, 2022. 1

[31] Shaoru Wang, Jin Gao, Zeming Li, Xiaoqin Zhang, and Weiming Hu. A closer look at self-supervised lightweight vision transformers. In *International Conference on Machine Learning (ICML)*. PMLR, 2023. 2, 3

[32] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[33] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2023. 1

[34] 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 (CVPR)*, 2022. 1, 7

[35] Sukmin Yun, Hankook Lee, Jaehyung Kim, and Jinwoo Shin. Patch-level representation learning for self-supervised vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[36] 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 (ICML)*, pages 12310–12320, 2021. 1, 2, 4, 5

[37] Qi Zhang, Yifei Wang, and Yisen Wang. How mask matters: Towards theoretical understandings of masked autoencoders. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3, 4, 5, 7, 8

[38] Zhiyu Zhao, Bingkun Huang, Sen Xing, Gangshan Wu, Yu Qiao, and Limin Wang. Asymmetric masked distillation for pre-training small foundation models. *arXiv preprint arXiv:2311.03149*, 2023. 2