2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

CMAE:Channel-Masked Autoencoders

Yutao Wang[1], Yu Nie[2], and Yilai Zhang[3]

[1] Jingdezhen Ceramic University, Jingdezhen City Jiangxi Province 333403, China
2220044008@stu.jcu.edu.cn
[2] Jiangxi Province Engineering and Technology Research Center of Ceramic Enterprise Informatization, Jingdezhen City Jiangxi Province 333403, China
nieyu@jcu.edu.cn
[3] Jiangxi Province Engineering and Technology Research Center of Ceramic Enterprise Informatization, Jingdezhen City Jiangxi Province 333403, China
zhangyilai@jcu.edu.cn

Abstract. In the field of self-supervised learning, representative frameworks such as MAE (Masked Autoencoder) and MoCo (Momentum Contrast) have emerged successively. However, the inherent architectural dependencies and implementation complexities of these methods constrain their universality and scalability across diverse model structures. Although existing training paradigms demonstrate certain optimization effects on lightweight model performance, systematic research in this domain remains insufficient. Thus, this study proposes Channel-Masked Autoencoders (CMAE) through improvements to the classical MAE framework, aiming to explore novel pathways for enhancing lightweight model performance. CMAE effectively resolves the architectural compatibility issues between MAE and convolutional neural networks while further investigating the operational mechanisms of noise injection strategies on model performance. This methodology establishes a dual-stage learning framework: the encoder extracts latent representations through channel masking and dual-cropped grayscale images, while the decoder reconstructs the original images accordingly. This design is inspired by the cognitive mechanism that the human visual system exhibits higher sensitivity to textural and morphological features compared to color attributes. Furthermore, this research innovatively introduces a controllable salt-and-pepper noise injection strategy that disrupts local pixel spatial correlations, guiding the encoder to learn more robust feature representations, achieves an absolute accuracy improvement of 0.3% on the Mini-ImageNet benchmark. To comprehensively evaluate model performance, we conduct systematic experiments across multiple heterogeneous datasets and cross-domain tasks. Notably, MobileViTv3 (with 4.7M parameters) improves classification accuracy from the baseline 70.5% to 74.2% on the Mini-ImageNet dataset, surpassing MoCoV3 by 1.9%. These results fully validate the technical advantages of our method in lightweight model optimization.

Keywords: Computer Vision, Self-Supervised Learning, Visual Feature Learning, Transfer Learning.

1 Introduction

In the field of self-supervised learning, prominent frameworks such as Masked Autoencoders (MAE) [3] and MoCo [15] have fully leveraged the potential of Vision Transformers (ViT) [4], demonstrating exceptional performance across diverse image processing applications. Notably, MAE has rapidly emerged as a mainstream approach for visual representation learning due to its robust learning capacity and scalability. However, a prevalent issue persists: many self-supervised methods are explicitly designed for predetermined architectures. For instance, MAE is specifically tailored for ViT and remains incompatible with convolutional neural networks (CNNs) [34] that employ sliding window operations. To address these architectural limitations, several improvement strategies have been proposed in academia. The FCMAE framework introduced in ConvNeXt v2 [35] achieves masked pre-training for CNNs by replacing standard convolutions with sparse convolutions. However, experimental evidence indicates that such structural modifications degrade model generalization in dense prediction tasks, and their compatibility with other architectures requires further validation. In contrast, the CMAE method proposed in this work realizes effective masked learning while preserving the integrity of native architectures.

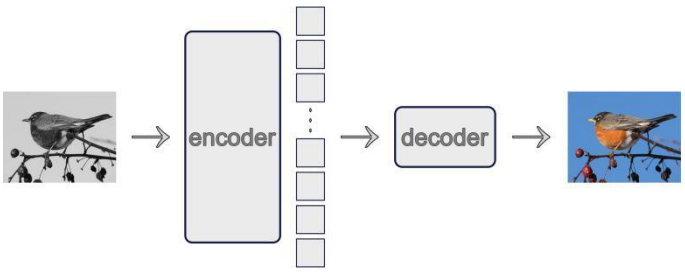

Fig. 1: The streamlined design of the CMAE framework.

Meanwhile, the rapid advancement of large-scale deep learning models has demonstrated remarkable performance on diversified datasets, primarily attributed to their parameter-intensive architectures [36–38]. Nevertheless, their computational intensity poses significant challenges in resource-constrained environments such as embedded and mobile systems, particularly regarding model training, optimization, and deployment. Although lightweight architectures have emerged to enhance computational efficiency, existing solutions generally face performance trade-offs, and research on deep optimization for resource-constrained scenarios remains notably deficient. Within the current computer vision ecosystem, the implementation of efficient network models is paramount. It is noteworthy that research in lightweight models predominantly focuses on architectural innovations [13, 39, 40, 25], while systematic investigations into training strategy optimization remain limited. We posit that training methodology optimization holds equivalent importance, especially considering the demonstrated efficacy of pre-training paradigms in large-model scenarios [23]. Of particular interest, conven-

tional self-supervised approaches like MoCov2, though successful in large architectures, underperform in lightweight models [24]. This underscores the critical need for customized pre-training techniques targeting resource-efficient models.

Noise mechanisms are conventionally perceived as interference factors in deep learning research. However, recent theoretical investigations reveal their dual nature as positive and negative noise [6]. Through comparative experimental analyses of CNN and ViT architectures, studies demonstrate that positive noise can effectively enhance the representational capacity of deep learning models. Nevertheless, existing research lacks further exploration of noise injection's compatibility with visual representation learning and lightweight architectures, particularly the operational mechanisms within masked learning frameworks. Therefore, investigating the effectiveness of noise injection strategies in lightweight models and their applicability in pre-training constitutes significant research value.

This study proposes a novel pre-training paradigm specifically optimized for lightweight models. The theoretical foundation originates from a crucial observation of human visual cognition mechanisms: object recognition in real world scenarios predominantly relies on textural and morphological features rather than chromatic information. To validate this hypothesis, through RGB channel-selective masking (as illustrated in Fig. 2), we discover that single channel grayscale images can effectively preserve the textural structures and morphological features of target objects even under dual-channel information loss conditions, with each channel exhibiting distinct visual representations. This finding theoretically aligns with existing conclusions in literature [7, 8] regarding the dominant role of textural-morphological features in specific visual tasks.

Fig. 2: Two color channels were removed, preserving only R, G, or B grayscale images.

Building upon this foundation, this paper constructs an efficient, concise, and universal Channel-Masked Autoencoder (CMAE) for visual representation learning. Compared to classical MAE, CMAE achieves deep compatibility with CNN architectures while maintaining Vision Transformer compatibility through an innovative channel-level masking mechanism: input images undergo random channel masking and dual random cropping before the encoder extracts latent representations, with a lightweight decoder subsequently reconstructing images based on these representations.

This study encompasses heterogeneous datasets for image classification and object detection. By constructing a multi-model comparative validation framework, we systematically analyze the scenario adaptability characteristics of the CMAE algorithm. Experimental results demonstrate that CMAE exhibits robust performance advantages across various scenarios: Compared with supervised learning methods, it achieves up to 3.4% improvement in classification accuracy and a 1.4% gain in detection mAP metrics. When benchmarked against MoCov3, the classification accuracy improvement reaches 1.9%. Notably, the additionally introduced noise injection strategy effectively

drives model performance enhancement, yielding an absolute improvement of 0.3%, thereby further validating the scalability of CMAE. The principal contributions are summarized as follows:

1. This study systematically investigates pre-training methodologies for lightweight architectures, with a focused exploration on CNN-ViT hybrid model frameworks. To our knowledge, this work represents the first systematic investigation in this research domain.

2. We propose the Channel-Masked Autoencoder (CMAE), a novel pretraining paradigm that effectively addresses the inherent compatibility challenges between MAE frameworks and convolutional networks, enabling adaptation to diverse visual architectures for pre-training requirements. Through systematic experimental validation, CMAE demonstrates marked performance superiority and exceptional scalability, thereby paving a new pathway for visual model pretraining.

3. Building upon the elucidation of noise mechanisms' impact on lightweight model performance, we innovatively propose a pixel-level noise injection strategy. Experimental data demonstrate that this strategy consistently enhances model accuracy, achieving an absolute performance gain of 0.3 percentage points.

2 Related Work

Self-supervised Learning predominantly focuses on diverse pretext tasks for images or texts [1, 9, 10]. Among these, contrastive learning [11] has gained significant popularity, exemplified by methods such as [12, 13], whose core mechanism lies in modeling similarity metrics and dissimilarity constraints (or single dimensional similarity optimization [14]) between multi-view data representations. Contrastive learning and related methodologies exhibit heavy reliance on data augmentation [13, 14]. A particularly noteworthy approach is MoCoV3 [15], a contrastive learning-based self-supervised framework capable of pre-training both CNN and ViT models. However, its essential components—momentum encoders and negative sample dependencies—result in framework complexity and high computational costs. In contrast, the proposed CMAE framework demonstrates relative simplicity and accelerated training efficiency through masking mechanisms.

Autoencoders represent classical paradigms in representation learning theory, comprising bidirectional mapping components: an encoder implementing nonlinear projection from input space to latent space, and a decoder reconstructing original inputs from latent representations. Classical algorithms with autoencoder architectures include Principal Component Analysis (PCA) and k-means clustering [16]. Denoising Autoencoders (DAE) [17], as critical extensions, achieve robust representations through "corruption-reconstruction" learning mechanisms. Existing studies indicate that numerous self-supervised methods can be viewed as instantiations of generalized DAE frameworks under various data perturbation conditions, with typical perturbation strategies including pixel masking [2, 18] and color image inpainting [10]. CMAE fundamentally differs from colorization tasks: while the latter primarily aims to generate realistic chromatic information from grayscale images, CMAE operates as a denoising autoencoder

that requires authentic reconstruction of original textural features beyond mere color recovery.

Masked Autoencoder (MAE) constitutes a Vision Transformer (ViT)-based self-supervised paradigm that achieves deep feature representation learning through an "encoding-reconstruction" pre-training mechanism. Its core architecture employs asymmetric encoder-decoder design: the encoder processes only unmasked visible patches, while the decoder executes pixel-level reconstruction using trainable mask tokens. This block-wise masking-based pre-training strategy, recognized for architectural simplicity and training efficiency, has been validated as an effective solution for constructing scalable visual representation systems.

Nevertheless, inherent limitations emerge when applying MAE frameworks to CNN pre-training, primarily stemming from fundamental architectural disparities between Transformers and CNNs. Specifically: (1) ViT's global self-attention mechanism inherently accommodates block-wise masking, with positional embeddings explicitly modeling spatial relationships of masked regions, whereas CNN's local convolution operations struggle to capture spatial correlations among discontinuous masked areas; (2) Weight-sharing properties of standard convolution kernels prevent effective differentiation between masked and valid feature regions, introducing noisy gradients during back propagation; (3) Intrinsic translation-equivariance in CNNs fundamentally conflicts with spatial position-awareness required by MAE frameworks. Preliminary experiments confirm significant performance degradation when training CNNs with simple block masking. Tian et al. [19] further attribute this performance gap to CNN's lack of explicit positional encoding mechanisms, hindering accurate spatial-semantic recovery of masked regions.

To address these architectural constraints, academic efforts have proposed multiple improvements. The FCMAE framework in ConvNeXt v2 achieves masked CNN pre-training by substituting standard convolutions with sparse convolutions. This approach suppresses gradient propagation in invalid masked regions through dynamic mask-aware convolutional kernel sparsification. However, empirical evidence reveals that such structural modifications impair model generalization in dense prediction tasks, with cross-architecture compatibility requiring further validation. Comparatively, the proposed CMAE introduces an innovative channel masking mechanism, establishing joint channel-spatial masking strategies to enable effective feature learning while preserving native architectural integrity.

Recent advancements in MAE frameworks demonstrate multi-dimensional innovations. Gao's team [19] significantly enhanced fine-grained feature capture through multi-scale encoding mechanisms, while Fei et al. [20] integrated Generative Adversarial Networks (GANs) to construct hybrid pre-training frameworks with competitive feature generation capabilities. These improvements validate MAE's extensibility advantages. Building upon these foundations, CMAE introduces three key enhancements: (1) Channel masking establishes robust representations through random color channel suppression; (2) Dual-cropping technology strengthens collaborative modeling of local-global features; (3) Ultra-low-rate noise injection improves model adaptability to distribution-shift scenarios.

Lightweight Models primarily address deployment requirements in resource constrained environments like edge devices and embedded systems, prioritizing balance between power consumption, latency, and performance. Beyond architectural optimizations, performance enhancement strategies focus on two technical directions: (1) Network topology optimization through parameter compression techniques like depthwise separable convolutions and channel pruning; (2) Transfer learning paradigm innovations, where knowledge distillation [21, 22] achieves knowledge transfer from complex to lightweight architectures via feature distribution alignment. Although pre-training techniques show potential for lightweight model optimization, systematic research remains scarce, particularly regarding pre-training methodologies for lightweight CNNs. ViT-focused studies are similarly limited, with most efforts concentrated on knowledge distillation rather than pre-training techniques, as evidenced by [23, 24]. Recent work by Shaoru W and Jin G [23] reports MAE and MoCoV3 applications on lightweight ViTs. While MAE outperforms MoCoV3 when pre-trained and fine-tuned on ImageNet-1k, it exhibits incompatibility with numerous hybrid ViT-CNN architectures.

Hybrid ViT-CNN architectures have garnered substantial attention for combining attention mechanisms and convolutional advantages to create lightweight models surpassing pure CNNs and ViTs. For instance, Astroformer [26] achieves superior performance on small-scale datasets like CIFAR-100 [27]. We pre-trained Astroformer models of varying scales using CMAE. Additionally, MobileViT [28] and EdgeNet [29], as representative hybrid models, deliver competitive results on large datasets (e.g., ImageNet-1k [30]) and excel in downstream tasks. CMAE pre-training further enhances their performance, demonstrating framework versatility across architectures and tasks.

Noise Injection Strategies recently proposed by Xiaowei Yu et al. [6], represent an innovative approach. Although noise is conventionally perceived as detrimental in deep learning, their extensive experiments on CNNs and ViTs reveal its dual nature: positive and negative noise. Positive noise specifically demonstrates performance-enhancing effects. Typical noise modalities are categorized as: (1) Gaussian noise, characterized by random perturbations in observation or latent spaces with normal-distributed amplitudes [31]; (2) Linear transformation noise, involving affine transformations of raw images or latent representations with row equivalent transformation matrices [32]; (3) Salt-and-pepper noise, an image distortion phenomenon caused by random black/white pixel injections at image or latent representation levels [33]. Notably, linear transformation noise has been identified as positive noise that effectively improves model performance. This work further investigates noise effects in masked learning, ultimately proposing innovative ultra-low-rate noise injection methodologies.

3 Approach

The Channel-Masked Autoencoder (CMAE) proposed in this work adheres to an encoder-decoder architectural paradigm analogous to MAE. The encoding module specializes in the nonlinear mapping of partially observable signals to latent space representations, while the decoding module accomplishes high-fidelity reconstruction of the original signal distribution. Distinct from MAE's block-wise masking mechanism, CMAE innovatively employs a channel-level masking strategy, eschewing the input image patching approach and directly processing full resolution images as atomic units. For three-channel RGB inputs, the method constructs partially observable data through random masking of two color channels, which are subsequently fed into the encoder for feature mapping—a process compatible with diverse computer vision architectures, including CNNs. These extracted features are then input to the decoder to reconstruct the masked channels. The core mechanism is visually elucidated in Fig. 1, the following text will introduce its components one by one.

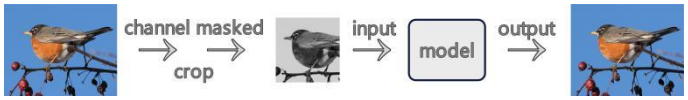

Fig. 3: Complete workflow of CMAE: from masking to reconstruction.

Masking. In the CMAE framework, we employ a random channel masking mechanism as the core strategy. The standard implementation constructs training samples by randomly masking two color channels. Since masking different channel combinations generates visually distinct single-channel images, this stochasticity inherently achieves a data augmentation-like effect. For pure Vision Transformer encoder architectures, this method supports the combined application of channel masking and random block masking.

However, training paradigms relying solely on random channel masking exhibit limitations in performance improvement, primarily due to inefficient optimization caused by global coverage of image regions in each training cycle. To address this, we innovatively integrate a dual random cropping strategy with channel masking: initial random cropping is applied during data pre-processing, followed by secondary random cropping before model input. This compound strategy functionally emulates block-wise masking. Images processed through dual random cropping serve as training inputs. Experimental validation demonstrates optimal acceleration benefits with a 50% cropping ratio while maintaining classification accuracy, as detailed in Fig. 3.

CMAE Encoder. In the CMAE design, the encoder processes cropped single channel grayscale inputs and maps them to latent space representations. Notably, this encoder is compatible with any native computer vision model architecture without structural modifications. This study primarily adopts lightweight hybrid architectures combining CNN and ViT advantages as the base encoder.

CMAE Decoder. For the decoder module, prior studies such as MAE and SimMIM [41] have demonstrated low correlation between decoder structure and model fine-tuning performance. Accordingly, CMAE employs a lightweight decoder only comprising four convolutional layers, inspired by the generator module in Generative Adversarial Networks (GANs) [42]. This design ensures efficient decoding while introducing negligible computational overhead.

The architecture details are shown in Fig. 4. The upsampling factors and channel dimension C in the decoder architecture require dynamic adjustment according to the input image resolution and the encoder's output feature dimensions. The current default parameter configuration (upsampling factor sequence 4→2→2) is designed for standard input resolutions of 224×224 pixels, based on the 7×7 feature map size generated through 32× downsampling in the encoder.

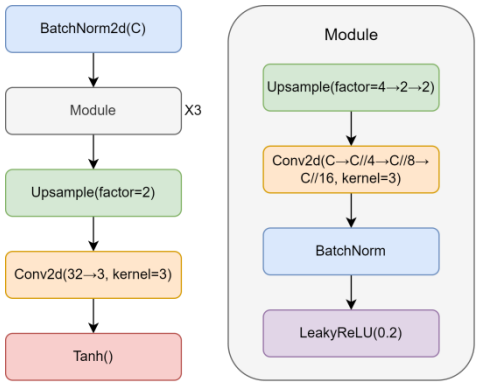

Fig. 4: CMAE Decoder.

Reconstruction Target. Regarding reconstruction objectives, both CMAE and MAE optimize for high-precision reconstruction of masked pixels, maintaining theoretical consistency in loss function design.

Fine-tuning. During fine-tuning, inputs are standard RGB three-channel images. To adapt to this input format, we implement a parameter expansion strategy: replicating single-channel pre-trained weights into a three-channel structure. This effectively resolves compatibility issues when transitioning from single channel pre-trained models to three-channel inputs, ensuring robust extraction of multi-channel visual features during fine-tuning.

Noise Injection Strategy. Two distinct noise injection techniques are adopted: an original pixel-level noise addition strategy and the Noisynn [6] method. Specifically, our proposed strategy injects 1%–5% low-intensity salt-and-pepper noise at the encoder input to simulate real-world image distortion while preserving feature learning stability. In contrast, the Noisynn [6] approach introduces 30% linear transformation noise before the encoder's final layers.

Compared to traditional self-supervised frameworks, CMAE's core innovation lies in its liberation from rigid architectural constraints. This architecture agnostic nature enables flexible adaptation to diverse model architectures, effectively resolving compatibility bottlenecks between MAE and CNNs. Whether applied to sliding-window convolutional operations or other computational paradigms, CMAE demonstrates exceptional architectural integration capabilities. Furthermore, the framework simplifies data flow by eliminating complex token embedding mechanisms and significantly reducing reliance on data augmentation, ultimately achieving orders-of-magnitude reductions in marginal computational costs during training.

In summary, CMAE provides a novel methodological perspective for self-supervised visual representation learning. Its breakthrough architectural universality and efficient engineering implementation not only enable seamless cross architecture integration but also reduce training resource consumption while preserving performance advantages. This framework offers new insights and directions for future research in visual representation learning.

4 Experiments

Our decision to avoid large-scale datasets for experimentation is driven by two principal considerations. First, while prior research predominantly focuses on large datasets—widely acknowledged for learning generalizable representations—the performance of models on small datasets remains an under explored yet noteworthy domain. Second, although Vision Transformers (ViTs) demonstrate superior performance over Convolutional Neural Networks (CNNs) under sufficient data conditions, CNNs retain irreplaceable advantages in few-shot learning scenarios. Consequently, this study employs small-scale benchmark datasets such as CIFAR-100 and Mini-ImageNet for experimental validation, with detailed performance metrics presented in Table 1.

This study deviates from conventional single-paradigm architectures like ViT or CNN. Current research in lightweight models prioritizes hybrid architectures that integrate convolutional operations with attention mechanisms, which exhibit significant advantages in energy efficiency ratios and inference latency, representing the technological frontier in computer vision. Aligning with this trend, we adopt convolution-attention hybrid architectures as base models, ensuring both alignment with domain evolution and technological innovation value.

A comprehensive explanation of parameter settings is provided in appendix. Notably, all experiments exclude external data sources, relying solely on standard augmentation techniques such as random cropping and horizontal flipping. Full training outcomes are derived via parameter fine-tuning, with strategies strictly adhering to original publications to ensure comparative fairness. Through a multi-dimensional evaluation framework, this study systematically validates the universal applicability of the CMAE pre-training strategy across heterogeneous datasets and diverse model architectures.

Table 1: Image classification experiments.

Datasets	Method	Params(M)	Supervised	CMAE
Cifar100	Astroformer	2.8	69.0	71.0+2.0
	MobileViTv3	4.7	64.7	68.1+3.4
ImageNet100	Edgexxsmall	1.3	67.0	71.4+4.4
	MobileViTv3	2.7	76.6	77.8+1.2
Mini-ImageNet	MobileViTv3	1.2	69.2	71.2+2.0
	MobileViTv3	4.7	70.5	74.2+3.7

4.1 Image Classification Experiments

This section presents a comparative evaluation of classification performance between CMAE-based pre-training/fine-tuning paradigms and training-from-scratch approaches across benchmark datasets, using Top-1 accuracy as the core metric.

Experiments focus on three datasets: CIFAR-100 [27], ImageNet-100, and Mini-ImageNet [43]. CIFAR-100, a standard dataset in computer vision, contains 60,000 32×32 RGB images spanning 100 fine-grained categories. ImageNet100 and Mini-ImageNet are subsets of ImageNet-1K [30]. The former comprises 100,000 color images across 100 classes, while the latter includes 60,000 images over 100 classes. Compared to ImageNet-100, Mini-ImageNet's reduced scale makes it suitable for model training and evaluation in resource-constrained environments.

Table 1 comprehensively compares performance across model architectures on diverse datasets. Crucially, experiments utilize the base CMAE framework without integrated noise modules to accurately assess core mechanism efficacy. Notably, results exclude CMAE performance on classical backbone networks like ResNet [44] and ViT. This omission arises not from framework incompatibility but from significant performance disparities between traditional architectures (ResNet, ViT) and hybrid models under identical parameter configurations and small-scale dataset conditions.

4.2 Comparative with Self-supervised Learning

In the comparative study of self-supervised learning paradigms, this work selects MoCov3 as the benchmark reference due to its representation of the technical frontier in general self-supervised frameworks. While other prominent methods like MAE and BEIT [36] hold theoretical innovation value, their exclusion from the comparison stems from compatibility constraints—specifically, their optimization for ViT architectures, which conflicts with the ViT-CNN hybrid architectures central to this study.

Contrastive learning frameworks typically rely on complex data augmentation strategies, and their multi-stage architectural designs incur significant computational resource demands. Despite this, the CMAE framework demonstrates superior performance over MoCoV3 across diverse scenarios, exhibiting varying degrees of advantage. As shown in Tables 2, we conduct comparative experiments on multiple

benchmark datasets, pre-training MobileViT-v3 models of varying scales using both MoCoV3 and CMAE. Results confirm that CMAE consistently outperforms MoCoV3.

Table 2: Comparative with self-supervised learning.

Datasets	Params(M)	MoCov3	CMAE
Cifar100	2.7	66.2	66.5+0.3
	4.7	67.6	68.1+0.5
Mini-ImageNet	4.7	72.3	74.2+1.9

4.3 Transfer Learning Experiments

To comprehensively evaluate transfer ability, we conduct object detection experiments on the PascalVOC dataset [45], comparing CMAE with supervised learning and MoCoV3 using mAP (mean Average Precision) metrics. All experiments are performed under strict closed-set conditions without external data. Results demonstrate CMAE's significant advantages in downstream task adaptability.

The PascalVOC dataset is widely recognized for its rich content and precise annotations, encompassing 21 object categories. All images include pixel level segmentation masks, bounding boxes, and class labels, providing a critical benchmark for diverse computer vision tasks. For object detection, the benchmark comprises 5,011 training and 4,952 test samples. The model architecture is based on MobileViTv3, serving as the backbone feature extractor for the single stage detector SSD [46]. Architectural modifications include replacing standard convolutions with depthwise separable convolutions in the SSD head, forming an SSDLite variant. Owing to its exceptional performance and lightweight nature, SSDLite is widely adopted for evaluating detection tasks in resource-efficient networks.

Table 3: Transfer learning experiments.

Method	supervised	MoCov3	CMAE
MobileViTv3-1.2M	56.8	57.8+1.0	58.0+1.2
MobileViTv3-2.7M	58.8	60.0+1.2	60.2+1.4

Table 3 systematically presents quantitative results from PascalVOC experiments. When scaling model capacity by increasing depth while fixing MobileViTv3's width parameter, CMAE's pre-training strategy consistently surpasses MoCoV3 across model sizes. Although performance gains fluctuate with scenario-specific conditions, CMAE maintains a significant advantage, validating the method's advanced scalability in model optimization.

4.4 Ablation Experiments

We conducted a series of ablation experiments to investigate the impact of varying parameter configurations on CMAE's performance during lightweight model pre-training.

These experiments revealed critical insights into model optimization and understanding. Notably, a theoretically simplified implementation path exists: applying zero-padding or mean-padding to masked image patches when using CNNs within the MAE framework. However, this approach was excluded from experimental results due to preliminary trials showing unsatisfactory performance—a finding consistent with theoretical expectations, as CNN's sliding window computation inherently struggles to extract generalized representations from locally masked regions.

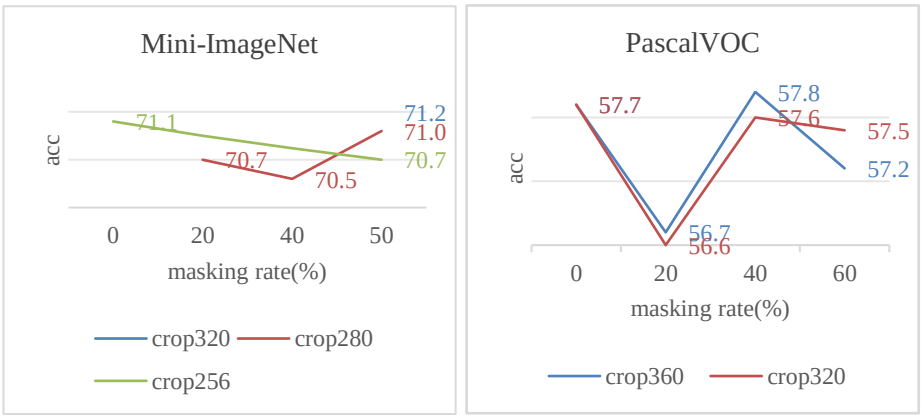

Fig. 5: Masking rate experiments.

Masking Rate. CMAE employs a dual-stage masking strategy: channel-level masking in the first stage and spatial-domain masking via secondary cropping in the second stage. This mechanism fundamentally differs from MAE's block level masking—CMAE focuses on masking redundant pixels outside cropped regions, prioritizing structural integrity of input images. Due to the synergistic effects of dual-stage masking, the overall masking rate may exhibit non-integer characteristics. When the encoder's downsampling factor is 32, the secondary crop size must satisfy two constraints: 1) The crop size must be an integer multiple of 32; 2) The downsampled size must be divisible by the initial crop size. Taking an initial crop size of 320 as an example, 160 can be selected as the secondary crop size (corresponding to a downsampled size of 5), which results in a calculated masking ratio of 50%. However, 192 cannot be chosen as the secondary crop size because its downsampled size becomes 6, and the initial crop size 320 is not divisible by 6. Leading to a dimensional mismatch in the feature maps, thereby rendering the decoder incapable of restoring the original dimensions.

Fig. 5 presents ablation results on masking rates using the MobileViTv31.2M model. We systematically explored the efficacy of block masking strategies across resolutions on the Mini-ImageNet and PascalVOC datasets. Experimental data demonstrate that a masking rate of ~50% (input resolutions of 160 or 192) achieves significant acceleration benefits while maintaining high accuracy. Further analysis reveals that increasing the initial crop size positively enhances model performance. However, compared to MAE's robustness under high masking rates, CMAE exhibits more pronounced perfor-

mance degradation in similar scenarios. This discrepancy arises from architectural differences between convolutional operations and self-attention mechanisms in feature extraction, particularly CNN's heightened sensitivity to local structural information.

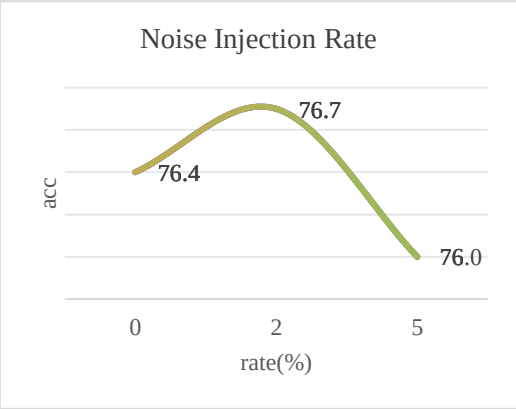

Fig. 6: Noise injection rate (salt-and-pepper noise).

Noise Injection Rate. Inspired by advanced noise models such as Noisynn [6], this study proposes a pixel-level random zeroing interference mechanism for noise injection strategy design. The specific implementation involves randomly selecting input image pixels for zeroing operations. This strategy enhances model robustness against noise contamination and suppresses feature redundancy by introducing perturbations in the feature space. As illustrated in Figure 6, experimental validation on the ImageNet-100 dataset using the MobileViTv3-1.2M architecture demonstrates that this noise scheme achieves an absolute accuracy improvement of 0.3 percentage points, fully validating CMAE's technical advantages in architectural scalability and parameter space optimization. Parameter sensitivity analysis reveals optimal classification accuracy when the noise ratio is set to 1%.

Furthermore, this study comparatively evaluates the enhancement effects of linear transformation noise. As shown in Table 4, the performance of the CMAE model is systematically improved by incorporating the linear transformation noise injection paradigm proposed in Noisynn [6]. Notably, this strategy exhibits positive gain effects even in small-scale datasets and lightweight model architectures, confirming the universal applicability of noise enhancement strategies across computational units of varying scales.

Table 4: Linear transformation noise.

Datasets	Method	Original	Noisy
PascalVOC	CMAE	57.7	58.0+0.3
Cifar100	supervised	64.7	65.1+0.4

Training Schedule. In the investigation of pre-training schedule regulation, Figure 7 quantitatively reveals the impact of pre-training duration on finetuning performance. Notably, excessively prolonged pre-training cycles may induce model overfitting risks—a phenomenon fundamentally distinct from MAE framework behaviors. In CMAE, each training cycle ensures full coverage of global image regions (with only color channel masking applied). The standard pre-training protocol is set to 700 epochs, but experiments indicate that model performance converges to saturation after 500 epochs. Although CMAE and MAE maintain formal consistency in pre-training strategies, their training dynamics diverge significantly due to structural differences in masking mechanisms and encoder design. Specifically, CMAE exhibits a stronger propensity for overfitting during extended training cycles.

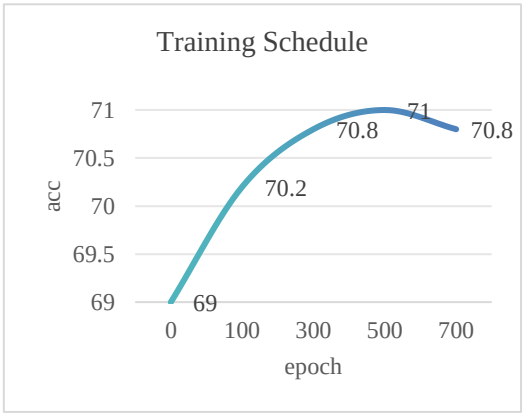

Fig. 7: Impact of training schedules.

5 Discussion and Conclusion

Conclusion. This study proposes the Channel-Masked Autoencoder (CMAE) with a dedicated focus on lightweight models and small-scale dataset investigations. By innovatively integrating channel-level random masking and dual spatial cropping mechanisms, CMAE effectively resolves the compatibility bottleneck between existing self-supervised frameworks (e.g., MAE) and specific architectures (e.g., CNNs). Through a lightweight decoder that reconstructs the original image distribution from latent representations, CMAE demonstrates exceptional architectural compatibility and domain adaptability, achieving a balance between performance enhancement and computational efficiency. Further introduction of noise intervention mechanisms systematically improves model robustness against adversarial perturbations and cross-domain stability. Empirical studies across datasets, tasks, and architectures paradigms substantiate the viability of CMAE as a general-purpose visual representation learning instrument.

Discussion. In the evolution of deep learning technologies, the scalability of algorithmic frameworks and engineering simplicity constitute core elements for sustaining a

healthy technical ecosystem. Highly scalable methodologies enable adaptation to rapid architectural iterations, while architecture-agnostic characteristics mitigate risks of technological path dependence, thereby extending algorithmic longevity. Building upon these principles, this study proposes the Channel-Masked Autoencoder (CMAE), which constructs an architecture agnostic visual representation framework through channel masking and dual stage cropping mechanisms. Its core innovation lies in translating the texture morphology perception mechanism of human vision into channel-level masking strategies, providing cross-scenario applicability for lightweight models.

Experimental validation demonstrates CMAE's effectiveness under low computational overhead. Nevertheless, two critical performance boundaries require further exploration: 1) Feature consistency is influenced by task diversity (e.g., fine-grained classification and open-domain detection), necessitating comprehensive multi-task evaluation metrics; 2) Its architecture-agnostic nature suggests potential scalability in large-scale models (e.g., Transformers). Future research should focus on two directions: analyzing generalization critical points through open-domain benchmarks and verifying adaptability in massive architectures.

References

1. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: ICLR (2018)
2. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
3. He, K., Chen, X., Xie, S., Li, Y., Dollar, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
5. Liu, Y., Zhang, Z., Chen, H.: Multi-Token Prediction with Hybrid Expert Networks: A Case Study on DeepSeek-V3. In: NeurIPS, pp. 1–15 (2024)
6. Yu, X., Huang, Z., Xue, Y., Zhang, L., Wang, L., Liu, T., Zhu, D.: Noisynn: Exploring the influence of information entropy change in learning systems. arXiv:2309.10625 [cs.AI] (2023)
7. Xie, Y., Richmond, D.: Pre-training on grayscale imagenet improves medical image classification. In: ECCV (2018)
8. Ge, Y., Xiao, Y., Xu, Z., Wang, X., Itti, L.: Contributions of shape, texture, and color in visual recognition. In: ECCV (2022)
9. Doersch, C., Gupta, A., Efros, A.A.: Unsupervised visual representation learning by context prediction. In: ICCV (2015)
10. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
11. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: CVPR (2006)
12. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020)
13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)

14. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR (2021)
15. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: ICCV (2021)
16. Hinton, G.E., Zemel, R.S.: Autoencoders, minimum description length, and helmholtz free energy. In: NeurIPS (1994)
17. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: ICML (2008)
18. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: ICML (2020)
19. Gao, P., Ma, T., Li, H., Lin, Z., Dai, J., Qiao, Y.: Convmae: Masked convolution meets masked autoencoders. arXiv:2205.03892 [cs.AI] (2022)
20. Fei, Z., Fan, M., Zhu, L., Huang, J., Wei, X., Wei, X.: Masked auto-encoders meet generative adversarial networks and beyond. In: CVPR (2023)
21. Bucilu ̧a, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 535–541 (2006)
22. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
23. Wang, S., Gao, J., Li, Z., Zhang, X., Hu, W.: A closer look at self-supervised lightweight vision transformers. In: ICML (2023)
24. Fang, Z., Wang, J., Wang, L., Zhang, L., Yang, Y., Liu, Z.: Seed: self-supervised distillation for visual representation. In: ICLR (2021)
25. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. In: NeurIPS (2021)
26. Dagli, R.: Astroformer: More data might not be all you need for classification. In: ICLR (2023)
27. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images (2009)
28. Wadekar, S.N., Chaurasia, A.: Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. arXiv:2209.15159 [cs.CV] (2022)
29. Maaz, M., Shaker, A., Cholakkal, H., Khan, S., Zamir, S.W., Anwer, R.M., Khan, F.S.: Edgenext: Efficiently amalgamated cnn-transformer architecture for mobile vision applications. In: ECCV (2022)
30. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009)
31. Russo, F.: A method for estimation and filtering of gaussian noise in images. In: IEEE Transactions on Instrumentation and Measurement, pp. 1148–1154 (2003)
32. Sun, T., Lu, C., Zhang, T., Ling, H.: Safe self-refinement for transformer-based domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7191–7200 (2022)
33. H.R., Ho C.W. Chan, Nikolova, M.: Salt-and-pepper noise removal by mediantype noise detectors and detail-preserving regularization. In: IEEE Transactions on Image Processing, pp. 1479–1485 (2005)
34. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. In: Neural Computation (1989)
35. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: CVPR (2023)

36. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. In: ICLR (2022)
37. Clark, K.L., Luong, M., Le, Q.V., Manning, C.D.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2019)
38. Zhai, X., Kolesnikov, A., Houlsby, N., Beyer, L.: Scaling vision transformers. In: CVPR (2022)
39. J. P., A. B., F. T., X. Z., L. D., H. L., G. T., B. M.: Edgevits: Competing lightweight cnns on mobile devices with vision transformers. arXiv:2205.03436 [cs.AI] (2022).
40. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR (2018)
41. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: CVPR (2022)
42. I. G., J. M., M. M., B. X., S. O., A. C., Y. B.: Generative adversarial networks. arXiv:1406.2661 [cs.AI] (2014).
43. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: NeurIPS (2016)
44. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
45. Everingham, M., Eslami, S.A., Parkhi, A.M., Torr, V.G., Williams, C., Reid, I.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. In: IJCV (2015)
46. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: ECCV (2016)