MaskAnyNet: Rethinking Masked Image Regions as Valuable Information in Supervised Learning

Jingshan Hong[1,2], Haigen Hu[1,2*], Huihuang Zhang[1,2], Qianwei Zhou[1,2], Li Zhao[3]

[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, Zhejiang, 310014, China
[2]Zhejiang Key Laboratory of Visual Information Intelligent Processing, Hangzhou, Zhejiang, 310014, China
[3]Zhejiang Normal University, Jinhua, Zhejiang, 321000, China
2112101525@zjut.edu.cn, hghu@zjut.edu.cn, huihuang@zjut.edu.cn, zqw@zjut.edu.cn, zhaoli2023@zjnu.edu.cn

Abstract

In supervised learning, traditional image masking faces two key issues: (i) discarded pixels are underutilized, leading to a loss of valuable contextual information; (ii) masking may remove small or critical features, especially in fine-grained tasks. In contrast, masked image modeling (MIM) has demonstrated that masked regions can be reconstructed from partial input, revealing that even incomplete data can exhibit strong contextual consistency with the original image. This highlights the potential of masked regions as sources of semantic diversity. Motivated by this, we revisit the image masking approach, proposing to treat masked content as auxiliary knowledge rather than ignored. Based on this, we propose MaskAnyNet, which combines masking with a relearning mechanism to exploit both visible and masked information. It can be easily extended to any model with an additional branch to jointly learn from the recomposed masked region. This approach leverages the semantic diversity of the masked regions to enrich features and preserve fine-grained details. Experiments on CNN and Transformer backbones show consistent gains across multiple benchmarks. Further analysis confirms that the proposed method improves semantic diversity through the reuse of masked content.

Introduction

Deep learning has been the cornerstone of progress in visual recognition tasks, driving remarkable advances in image classification, detection, and segmentation through large-scale annotated datasets and more advanced model architectures. (Yang et al. 2024; Brown et al. 2020). Image masking methods, as a common data augmentation approach, have been widely adopted in supervised learning and self-supervised image reconstruction (Shikhar and Sobti 2024; Feng and Zhang 2024). These methods improve model robustness by occluding parts of the image, forcing the model to make predictions or reconstructions based on incomplete information. Although these methods have shown effectiveness in regularization and self-supervised representation learning, in the field of supervised learning, they usually treat masked pixel information as useless noise rather than a reconfigurable and reusable source of information. This will lead to two critical limitations: (i) Direct discarding of masked region pixel fails to fully exploit available image information, thereby constraining the model's input utilization efficiency; (ii) Critical semantic elements (e.g. small objects and fine-grained textures) may be obscured, which inhibits feature learning in such regions and consequently compromises recognition capability for small or occluded targets.

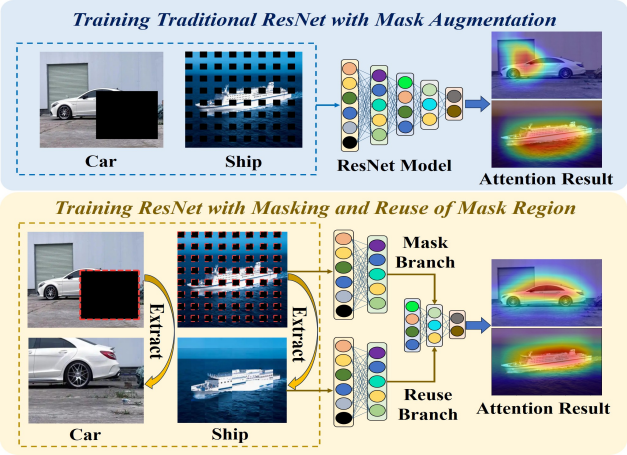


Figure 1: Comparison of effect between conventional mask discarding and with reuse strategies for complementary visual information based on ResNet-34. Top: the heatmaps can only cover part of the target and fails to clearly perceive edge details of the target. (e.g., partial coverage of vehicles or distraction by background elements like ocean scenes). Bottom: the heatmaps precisely localize the target by repurposing masked regions as complementary visual cues.

We note that masked image modeling (MIM) enables the reconstruction of comprehensive visual representations from minimal visible pixel subsets (Li et al. 2025; Wang et al. 2025), which means that even a limited set of unmasked patches can preserve the global structural semantics of the original image. Motivated by these observations, we rethink masked image regions in supervised learning: rather than discarding them as noise, we propose to repurpose these regions as complementary visual information sources. Specifically, we explore whether image masking can serve as a mechanism to provide diverse and fine-grained features, thus

*Corresponding author

enhancing feature representation learning.

Consequently, we conducted preliminary investigations along this direction. Figure 1 illustrates the comparative effects of reusing masked regions as complementary visual information sources by using class activation heatmaps. From Figure 1, the traditional masking approach fails to capture complete features and perceive the clear details of the edge, thus limiting the ability of the model to make accurate predictions. While the reuse strategy for complementary visual information can achieve precise target localization with full-contour awareness while effectively suppressing background noise, it means that the information reuse branch recovers these details through masked latent learning.

Based on these insights, we propose a dual-branch architecture, called MaskAnyNet, to repurpose discarded mask regions as complementary visual information sources. Specifically, we redesign backbone models into a dual-branch architecture that simultaneously learns from both visible content and masked regions. A primary branch processes the masked image to extract global features, while another auxiliary branch extracts fine-grained details from masked regions to refine feature representations. This integrated design leverages both the regularization benefits of masking and enhanced representational capacity from reintegrated local features, enabling comprehensive pixel-level information utilization and multi-perspective visual understanding. **The main contribution of this work can be summarized as follows**.

- **Architecture Innovation**: We propose **MaskAnyNet**, a novel unified architecture featuring a *mask-guided information reuse branch*. Compatible with both CNNs and Transformers, it consistently boosts baseline performance across diverse datasets while maintaining computational efficiency comparable to vanilla backbones.
- **Mask Pattern Analysis**: We rigorously characterize masking and reuse patterns, establishing generalized configuration guidelines. Using *information entropy* and *cross-region similarity metrics*, we quantitatively validate mask strategy effectiveness, providing theoretical grounding for our approach.

Related Work

Image Masked Data Augmentation

Image masking is a widely used method in the field of computer vision(Wei et al. 2017; Murdock et al. 2016; Wang, Shrivastava, and Gupta 2017; Ghiasi, Lin, and Le 2018). By masking specific areas of the input image, the model focuses more attention on the area of interest, thereby improving the generalization ability of the model. In the field of data augmentation, some typical methods such as CutOut(DeVries 2017) and Random Erasing(Zhong et al. 2020), randomly mask parts of the image to encourage the model to focus on different parts of the target, preventing over-reliance on certain local features and significantly improving the model's generalization ability. Building on this, Hide-and-Seek(Kumar Singh and Jae Lee 2017) employs a broader range of random mask patch to enable the model to learn features from more areas of the target, thus improving its recognition ability when the target is occluded or incomplete. Furthermore, GridMask(Chen et al. 2020) adopts a structured patch strategy by regularly generating grid-shaped mask, achieving a better balance between information removal and retention. FenceMask(Li, Li, and Long 2020) generates continuous mask areas in a fence-like shape, making it more suitable for processing scenes with small targets. It is highly effective in preserving small object features and can be better adapted to datasets with complex small target features.

Masked Image Modeling

MIM has emerged as a cornerstone of self-supervised learning, enabling both image restoration (Qin et al. 2024; Chen et al. 2024; Wei et al. 2022) and transferable visual representation learning (Woo et al. 2023; Dong et al. 2023). Inspired by NLP's Masked Language Modeling (MLM), MIM reconstructs occluded regions to learn annotation-free visual features (Zhan et al. 2023). BEIT (Bao et al. 2021) extends BERT's approach(Devlin 2018) by predicting representations of randomly masked patches, significantly enhancing feature quality. BEITv2 (Peng et al. 2022) further advances performance through joint optimization and semantic refinement. MAE(He et al. 2022) reconstructs full images from partial inputs to learn deep representations. SimMIM(Xie et al. 2022) unifies masked region prediction and representation learning in a single encoder, reducing model complexity. OmniMAE(Girdhar et al. 2023) extends MAE to multimodal joint modeling. PixMIM(Liu et al. 2023) directly reconstructs masked pixel values, eliminating reliance on pretrained tokenizers. MixMask(Vishniakov, Xing, and Shen 2022) uses a CutMix-based distance loss that computes similarity through mask ratios for precise image matching.

While masking excels in data augmentation and self-supervised learning, its potential in supervised settings remains limited by treating masked regions as disposable noise, causing valuable information loss. This motivates our fundamental reexamination of masked signal utilization within supervised paradigms.

Methodology

We propose MaskAnyNet, a novel architectural paradigm that integrates region masking with a reuse branch to reintroduce masked pixel information into the learning process. As shown in Figure 2, the framework comprises three components: image masking branch, mask region reuse branch, and feature fusion and alignment. First, the image masking branch occludes parts of the input using various masking strategies. Then, the reuse branch reintroduces the original content of masked regions to enhance contextual understanding. Finally, the fusion and alignment module integrates features from both branches to optimize their interaction. The details of each component are described below.

Image Masking Branch

The image masking methods can be classified into three types: single Patch mask, Random mask, and grid mask, as illustrated in Figure 3. As shown in the Figure 3, grid masking can be regarded as a uniform sampling strategy. When

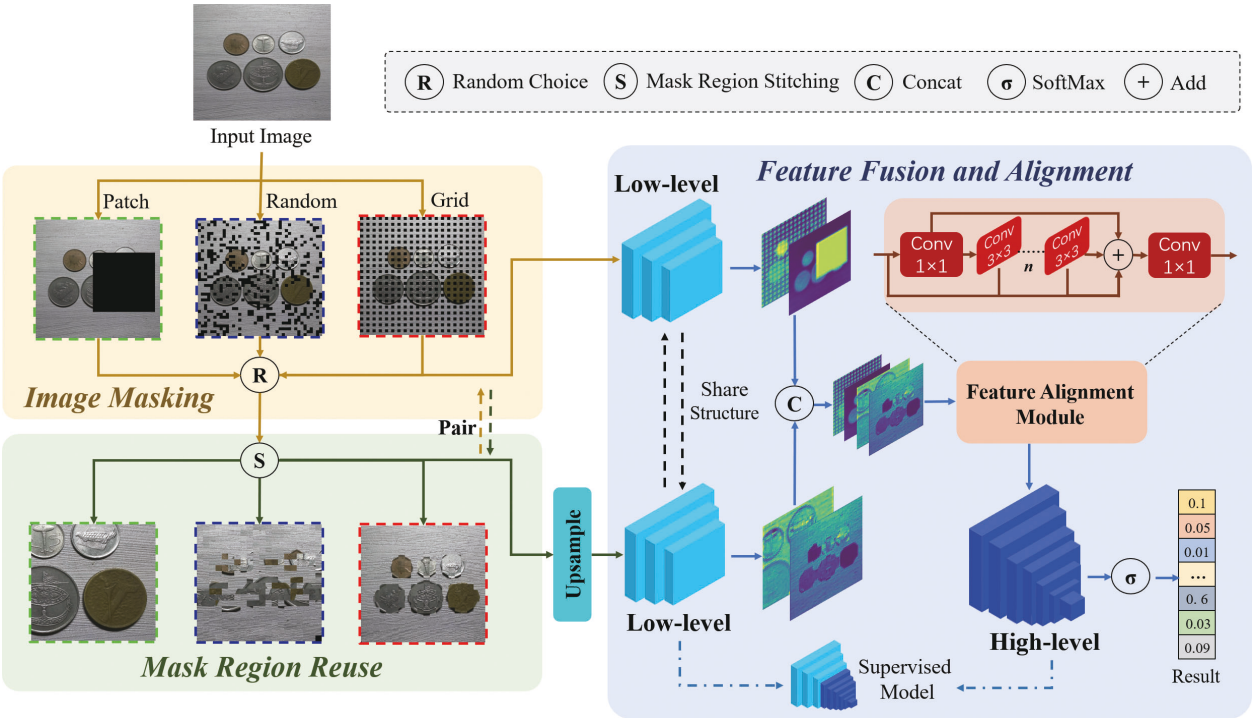
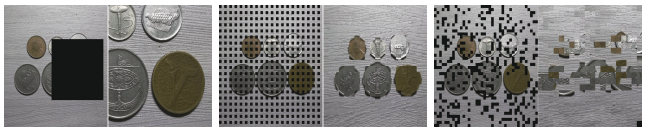


Figure 2: Overall architecture of the proposed MaskAnyNet, which consists of three main components: mask generation, mask-region information reuse, and feature fusion and alignment.



(a) Patch Mask (b) Grid Mask (c) Random Mask

Figure 3: Comparison of different masking methods. Patch Masking preserves local details, Grid Masking preserves global semantics through fixed-pattern, and Random Masking enhances diversity through irregular patterns.

the masked regions are stitched together, it helps preserve essential global structure. This allows the model to efficiently model multi-scale features and enhance the effective receptive field from limited visible pixels.The regular spatial coverage of grid masking also facilitates the extraction of long-range dependencies with minimal information loss. In contrast, the patch masking strategy emphasizes local detail. By randomly masking fine-grained patches, the model can learn diverse local features such as textures, edges, and small-scale object parts. This improves the model's sensitiViTy to fine details and boosts its ability to represent small or irregular targets. Random masking introduces high variability in the spatial distribution of masked regions, which encourages the model to reason over more complex and unpredictable patterns. However, due to its unstructured nature, random masking introduces significant spatial distortion, such as positional ambiguity and geometry shifts. Although this property benefits self-supervised tasks like masked image reconstruction, it may undermine performance in supervised tasks like classification or detection, where consistent semantic coherence and spatial alignment are crucial. Based on the above analysis, our masking strategy integrates both grid and random patch masking. By combining grid and patch masking, it strikes a balance between global semantic preservation and local detail enhancement, enhancing the ability of model to learn diverse visual patterns.

Mask Region Reuse Branch

Although the mask image can serve as a regularization, it can result in the loss of key features, particularly in small target areas. Furthermore, the effectiveness of masking strongly depends on the mask strategy and mask ratio, which may not be universally suitable for all tasks or image distributions. To address these limitations, we design a mask information reuse branch, which is designed to re-learn and enhance the information lost due to masking. Specifically, let M be the set of masked regions in the input image, where each masked block is denoted as M_i, with $i \in [1, 2, ..., N]$, where N is the total number of masked blocks. The masked blocks are indexed based on their position in the original image. After that, identify all the masked blocks M_i within the input image and extract their corresponding original patch Ui from the original image at the same spatial location as M_i.

After extracting the masked regions, we combine all extracted patches U_i according to their relative spatial positions in the original image. The reuse image R is then con-

structed by Eq.1:

$$R = \mathcal{S}\left(\{(U_i, P_i)\}_{i=1}^{N}\right) \tag{1}$$

where P_i denotes the spatial coordinates of the i-th masked block in the original image, and $\mathcal{S}(\cdot)$ is a spatial stitching operator that arranges each extracted patch U_i according to P_i, preserving their relative spatial arrangement as in the original image. This spatially-aware composition preserves relative positional consistency and complements the masked input by reintroducing critical visual information lost during masking. By reintroducing these semantically rich but commonly overlooked areas, the reuse branch not only complements the local representation, but also reduces the risk of information loss in standard masking-based learning.

Feature Fusion and Alignment

Due to the size mismatch between the reuse image and the masked image, the reuse image is first upsampled to match the spatial dimensions of the masked image before being further processed. Then both images are input into a low-level feature extractor that shares the structure. The extractor is constructed by selecting the shallow layers from a standard deep neural network backbone. These early layers retain high-resolution spatial information and are well-suited for capturing fine-grained local patterns, including edges, corners, and textures. The extracted features from both inputs are concatenated along the channel dimension to form a joint representation.

To address the semantic inconsistency between global and local features, we design a feature alignment module, as illustrated in Figure 2. This module performs soft alignment by applying a series (n, default n=3) of convolutional layers and preserves fine-grained details and local texture through a multi-level residual refinement mechanism.

$$F_{\text{out}} = \mathrm{Conv}_{1\times1}\left(\sum_{i=1}^{n} \mathrm{Conv}_{3\times3}^{(i)}\left(F_i\right) + \mathrm{Conv}_{1\times1}\left(F_{\text{in}}\right)\right) \tag{2}$$

where F_i denotes the output feature of the i-th convolution block, and $Conv_{k\times k}^{(i)}$ denotes the i-th convolution layer with kernel size $k \times k$. After alignment, the resulting feature is forwarded to a high-level feature extractor, which consists of the deeper layers of the same backbone network used in the low-level feature extractor. This design ensures hierarchical consistency while allowing the model to capture enriched semantic features across multiple scales. It is precisely this modular decomposition that enables our method to be compatible with any backbone architecture and provides significant flexibility in design. Additionally, constraining the mask reuse branch to operate in the low-level feature extraction stage allows efficient control of additional parameters and computational cost.

Experiments and Results

Implementation Details

All experiments were conducted on a computer with an NVIDIA RTX 4090 GPU. The equipped software runtime environment was also set up with python 3.9, PyTorch 2.1.1, CUDA 11.8 and cuDNN 8.6.0. All baseline models used in the experiments, as well as the corresponding MaskAnyNet variants, are trained with the same hyperparameters as their original baseline implementations. Specifically, we train for 300 epochs on CIFAR datasets and 200 epochs on ImageNet.

CIFAR-10 and CIFAR-100 Experiments

To evaluate the effectiveness and generalizability of our approach, we conducted comparative experiments on CIFAR-10 and CIFAR-100 using representative CNN and Transformer models. The comparison results are presented in Table 1. Specifically, we selected ResNet-34 and EfficientNet-V2 as CNN baseline models, with the former being a widely used architecture and the latter achieving the best performance among the CNN baselines. For Transformer-based baselines, we chose ViT and Swin, both of which are well-established and effective in vision tasks. As shown in Table 1, our method improves the accuracy across all models on both CIFAR-10 and CIFAR-100. For example, ResNet-34 and EfficientNet-V2 gain 1.56% and 1.03% on CIFAR-10, respectively, while ViT and Swin improve by 0.42% and 0.67%. On CIFAR-100, the gains are even more pronounced, with up to 1.78% on ResNet-34, 1.42% on EfficientNet-V2, and 0.97% on ViT, demonstrating the strong generalization ability of our method. The larger gains on the more challenging CIFAR-100 dataset further suggest that the proposed reuse strategy is particularly effective for fine-grained recognition tasks.

ImageNet Experiments

To further validate the universality and robustness of MaskAnyNet, we conducted experiments on the large-scale ImageNet-1K and Tiny-ImageNet datasets, as presented in Table 2. The results show consistent improvements across all baseline. Specifically, on ImageNet-1K, all baseline models achieve more than a 1.0% improvement in Top-1 accuracy when integrated with our approach. Notably, the Top-1 accuracy of ViT-B/16 and ResNet-34 increases significantly by 1.56% and 1.45%, respectively. Furthermore, the performance gains on Tiny-ImageNet are more pronounced. Specifically, our method improves the Top-1 accuracy of various models, including 2.08% on ResNet-34, 1.28% on EfficientNetV2, 1.68% on ViT, and 1.71% on Swin, demonstrating consistent gains across both CNN and Transformer architectures. Compared to ImageNet-1k, the limited amount of training data in Tiny-ImageNet causes the baselines to suffer from poor generalization. In contrast, our method not only provides a regularization effect, but also enhances pixel utilization by reusing masked information. This enables the model to learn more robust and diverse feature representations. As a result, our method significantly improves the performance of the baseline models.

Comparative results across multiple datasets confirm that our masked region reuse strategy generalizes well across different backbone architectures. By transforming discarded pixels into reusable learning signals, our method facilitates richer feature learning and improves the model's ability to recognize small, occluded, or semantically diverse targets.

$\mathcal{D}$	Method	Top-1 Acc
CIFAR-10	ResNet-34(He et al. 2016)	94.21
	EfficientNetV2-S(Tan and Le 2021)	94.72
	MobileNetV4-Small(Qin et al. 2025)	93.23
	GhostNet-V3(Liu et al. 2024)	94.12
	FasterNet-S(Chen et al. 2023)	91.95
	ConvNextV2-Tiny(Woo et al. 2023)	94.67
	Swin-Tiny(Liu et al. 2021)	96.56
	ViT-Base(Dosovitskiy 2020)	98.99
	MaskResNet-34	**(+1.56)95.77**
	MaskEfficientV2	**(+1.03)95.75**
	MaskViT	**(+0.42)99.41**
	MaskSwin	**(+0.67)97.23**
CIFAR-100	ResNet-34(He et al. 2016)	73.58
	EfficientNetV2-S(Tan and Le 2021)	76.71
	MobileNetV4-Small(Qin et al. 2025)	71.78
	GhostNet-V3(Liu et al. 2024)	74.09
	FasterNet-S(Chen et al. 2023)	70.31
	ConvNextV2-Tiny(Woo et al. 2023)	72.32
	Swin-Tiny(Liu et al. 2021)	90.46
	ViT-Base(Dosovitskiy 2020)	91.90
	MaskResNet-34	**(+1.78)75.36**
	MaskEfficientV2	**(+1.49)78.20**
	MaskViT	**(+0.97)92.87**
	MaskSwin	**(+0.77)91.23**

Table 1: Comparison of Top-1 Accuracy of MaskAnyNet and different models on CIFAR-10 and CIFAR-100 datasets.

$\mathcal{D}$	Method	Top-1 Acc	Top-5 Acc
ImageNet-1K	ResNet-34(He et al. 2016)	71.67	90.64
	EfficientNetV2-S(Tan and Le 2021)	74.73	91.25
	MobileNetV4-Small(Qin et al. 2025)	68.95	89.62
	GhostNet-V3(Liu et al. 2024)	69.58	89.31
	FasterNet-S(Chen et al. 2023)	70.68	90.88
	ConvNextV2-Tiny(Woo et al. 2023)	74.68	91.54
	Swin-T(Liu et al. 2021)	78.63	93.87
	ViT-B/16(Dosovitskiy 2020)	79.51	94.62
	MaskResNet-34	**(+1.45)73.12**	**(+1.19)91.83**
	MaskEfficientV2-S	**(+1.12)75.85**	**(+0.29)91.54**
	MaskViT-B/16	**(+1.56)81.07**	**(+0.76)95.38**
	MaskSwin-T	**(+1.32)79.95**	**(+0.61)94.48**
Tiny-ImageNet	ResNet-34(He et al. 2016)	60.27	80.98
	EfficientNetV2-S(Tan and Le 2021)	64.57	84.84
	MobileNetV4-Small(Qin et al. 2025)	60.57	82.83
	GhostNet-V3(Liu et al. 2024)	58.49	81.42
	FasterNet-S(Chen et al. 2023)	58.67	84.32
	ConvNextV2-Tiny(Woo et al. 2023)	59.05	79.04
	Swin-T(Liu et al. 2021)	82.32	93.61
	ViT-B(Dosovitskiy 2020)	85.83	95.45
	MaskResNet-34	**(+2.08)62.35**	**(+1.28)82.26**
	MaskEfficientV2	**(+1.98)66.55**	**(+0.11)84.95**
	MaskViT	**(+1.68)87.51**	**(+0.96)96.41**
	MaskSwin	**(+1.71)84.03**	**(+0.61)94.22**

Table 2: Comparison of Top-1 and Top-5 accuracy of MaskAnyNet and different models on ImageNet-1K and Tiny-ImageNet datasets.

Ablation Experiments

To investigate the contribution of each component in our framework, we conducted ablation studies on the ImageNet-1K dataset using both ResNet-34 and ViT-B/16 as baseline, as shown in Table 3. Specifically, $\mathcal{M}$ denotes the use of mask augmentation, $\mathcal{R}$ indicates the mask region reuse branch, and FFA refers to the feature fusion and alignment module. Introducing masking augmentation leads to certain improvements, demonstrating the regularization and local perturbation benefits of masking. Further incorporating the reuse branch results in the most significant improvement, with Top-1 accuracy gains of 0.85% for ResNet-34 and 1.06% for ViT-B/16, confirming the effectiveness of reintroducing masked region information. Finally, introducing FFA results in the best performance, highlighting that it effectively bridges the semantic gap between both branches and improves the integration of global-local features.

Masking Strategy Analysis

We evaluated three masking strategies on the ImageNet-1K dataset using ResNet-34, and the results are shown in Table 4. As observed in the table, random masking yields the lowest performance, due to significant structural deviations in the reuse image, which hinder feature alignment between the masked and reused images and affect the model's learning process. Grid masking performs better by preserving global semantics. The reuse image generated through grid sampling effectively restores global information. However, its overall feature distribution closely resembles that of the masked image, which may cause the model to exhibit "lazy" behavior, relying too heavily on redundant information, thus limiting further performance improvement. In contrast, patch masking achieves the highest accuracy by introducing local variability and preserving critical object details. It provides regularization benefits for the masked image and ensures that key features are retained through the reuse process. Moreover, this regional reuse strategy enables the model to capture fine-grained local patterns, enhancing representational capacity and generalization. Notably, combining patch and grid masking achieves the best

Method	$\mathcal{M}$	$\mathcal{R}$	FFA	Top-1 Acc	Top-5 Acc
ResNet-34	-	-	-	71.67	90.64
Patch Mask	✓	-	-	72.12	91.25
Grid Mask	✓	-	-	71.97	91.14
Random Mask	✓	-	-	71.88	91.02
MaskResNet-34	✓	✓	-	72.97	91.76
MaskResNet-34	✓	✓	✓	73.12	91.83
ViT-B/16	-	-	-	79.51	94.62
Patch Mask	✓	-	-	79.88	94.78
Grid Mask	✓	-	-	79.95	94.83
Random Mask	✓	-	-	79.86	94.67
MaskViT-Base	✓	✓	-	81.01	95.23
MaskViT-Base	✓	✓	✓	81.07	95.38

Table 3: Ablation experiment on the ImageNet-1k dataset.

Patch	Grid	Random	Top-1 Acc	Top-5 Acc
✓			72.89	92.03
	✓		72.72	91.98
		✓	72.23	91.73
✓	✓		**73.12**	**91.83**
✓	✓	✓	73.07	91.67

Table 4: Performance comparison on the ImageNet-1k using different mask and reuse methods based on ResNet-34.

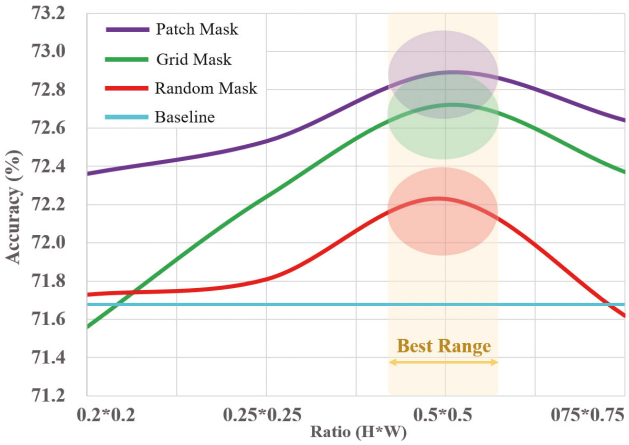


Figure 4: Performance visualization results of MaskResNet-34 using three different masking methods and different mask ratios on the ImageNet-1K dataset.

result, demonstrating the complementary effect of global structure and local detail reuse. Based on these findings, our final reuse strategy adopts both patch masks and grid masks.

Mask Ratio Analysis

In addition to the choice of masking strategy, the mask ratio also plays a crucial role in determining how much visual information is retained, thereby directly affecting the model's ability to learn diverse representations. To investigate this, we evaluated different masking ratios across different masking methods, and the results are presented in Figure 4.

As shown in Figure 4, a low mask ratio leads to weak regularization and insufficient information in the reused image, resulting in poor accuracy. As the ratio increases, the accuracy of all methods improves, reaching the best performance at a mask area ratio of 25%. Further increasing the ratio leads to a decrease in accuracy, indicating that excessive mask limits the model's ability to learn informative content.

Feature Fusion Strategy Analysis

To effectively integrate features from different branches, it is essential to choose an appropriate fusion location. The most common approaches include image-level fusion(Zhang et al. 2017; Zhou et al. 2016), feature-level fusion(Chen, Fan, and Panda 2021; Chi et al. 2019), and decision-level(Snoek, Worring, and Smeulders 2005; Baltrušaitis, Ahuja, and Morency 2018) fusion. We evaluate these three strategies on the ImageNet-1K using ResNet and pre-trained ViT-B/16. Table 5 shows the comparison of these three strategies in terms of accuracy and computational efficiency. It can be seen that the method of using feature fusion consistently achieves the highest accuracy in both ResNet and ViT, outperforming the other two methods. This proves that feature layer fusion can achieve good filtering of noise while aligning two branches information through remaining deep layers of the model, with high robustness.

Model	Fusion Level	Acc	Pa(M)	Latency(ms)
ResNet-34	Image-level	73.03	21.34	7.29
	Feature-level	73.12	22.46	7.51
	Decision-level	72.28	43.10	14.04
ViT-Base	Image-level	79.37	85.72	7.94
	Feature-level	79.51	87.46	8.23
	Decision-level	78.82	171.45	16.21

Table 5: Performance Comparison of Different Feature Fusion Strategies in Accuracy and Computational Efficiency.

In terms of computational efficiency, using low-level feature fusion can maintain a similar number of parameters as image-level fusion, and the computational burden introduced by the additional branch can be largely ignored. However, decision-level fusion nearly doubles the computational cost. Moreover, due to the depth of the fusion site, it becomes difficult to effectively align features from the two branches, resulting in limited improvement in accuracy.

Detection and Segmentation Experiments

Our method demonstrates notable performance gains on classification tasks, reflecting its strong feature representation capabilities. To further evaluate the generalizability of the proposed approach across downstream tasks, we conduct additional experiments on object detection and semantic segmentation, using representative CNN-based and Transformer-based models for comparison.

Object Detection

In the object detection task, we employed YOLO12(Jocher, Qiu, and Chaurasia 2023) (a CNN-based detection model) and RT-DETR(Zhao et al. 2024) (a Transformer-based detection model) as baseline models, testing on two authoritative datasets PASCAL VOC(Everingham et al. 2010) and MS COCO(Lin et al. 2014). Following (Girshick 2015), training and testing of the PASCAL VOC dataset were carried out by using 5k images from VOC 2007 and 16k images from VOC 2012 ("07+12") for training, while the VOC 2007 test set was utilized for evaluation.

We used P, R, mAP@.5, and mAP@[.5,.95] as the primary metrics, and the final comparison results are shown in Table 6. It can be seen that our method achieved accuracy gains in both YOLO12 and RT-DETR. Notably, YOLO12-n exhibited 1.68% and 1.89% improvements in mAP@.5 on VOC and COCO, respectively; meanwhile, RT-DETR showed 1.32% and 1.42% improvements under the same

Dataset	Model	Precision	Recall	mAP@.5	mAP@[.5,.95]
PASCAL	RT-DETR-n	82.46	75.52	81.63	62.53
	MaskRT-DETR-n	**(+0.57)83.03**	**(+0.76)76.28**	**(+1.32)82.95**	**(+0.85)63.38**
	YOLO12-n	80.29	74.18	81.19	61.69
	MaskYOLO12-n	**(+1.02)81.31**	**(+0.75)74.93**	**(+1.68)82.87**	**(+0.93)62.62**
MS-COCO	RT-DETR-n	68.27	53.93	63.83	44.82
	MaskRT-DETR-n	**(+0.92)69.19**	**(+0.74)54.67**	**(+1.42)65.25**	**(+1.04)45.86**
	YOLO12-n	65.51	49.58	54.46	38.86
	MaskYOLO12-n	**(+1.33)66.84**	**(+1.07)50.65**	**(+1.89)56.35**	**(+1.03)39.89**

Table 6: Object detection performance comparison of MaskAnyNet on MS COCO and PASCAL VOC datasets using RT-DETR and YOLO12 as baselines.

metrics. It is worth emphasizing that our method maintained robust detection performance for objects spanning a wide range of scales, indicating that the mask information reuse mechanism effectively enhances the cross-scale representational capacity of the model, thus alleviating the missed or false detections of targets with varying sizes.

Semantic Segmentation

For semantic segmentation task, we train DeepLabV3+ (Chen et al. 2018) (CNN-based) and Segformer (Xie et al. 2021) (Transformer-based) on ADE20K dataset(Zhou et al. 2019) to evaluate the performance improvements offered by our MaskAnyNet in high-granularity segmentation tasks. Both lightweight and deeper backbones were tested, and using mIoU as the primary evaluation metric, the results are shown in the Table 7. It can be observed that, regardless of the baseline method, all the models incorporating our approach achieve better performance. DeepLabV3+ improves its mIoU by 1.83% when using a lightweight backbone, while still achieving a gain of 1. 22% with the deeper ResNet-101 backbone. This phenomenon suggests that lightweight networks possess greater room for improvemen, an observation also supported by Segformer results, which achieves a 0.71% increase in mIoU on the deeper B4 backbone and a 1.28% increase on the lighter B0 backbone. Through both detection and segmentation experiments, it becomes clear that the proposed method demonstrates strong generalizability across diverse encoder architectures. In particular, it shows notable advantages on lightweight backbones, underscoring its effectiveness in balancing computational efficiency and accuracy.

Model	Encoder	mIoU
DeepLabV3+	MobileNetV2	34.00
	ResNet-101	44.10
MaskDeepLabV3+	MobileNetV2	**(+1.83)35.83**
	ResNet-101	**(+1.22)45.32**
Segformer	MiT-B0	37.40
	MiT-B4	51.10
MaskSegformer	MiT-B0	**(+1.28)38.68**
	MiT-B4	**(+0.71)51.81**

Table 7: Segmentation performance comparison of MaskAnyNet on ADE20K dataset using DeepLabV3+ and SegFormer as baselines.

Further Analysis

Image Similarity and Information Entropy

Although the performance improvement achieved by our method has been quantitatively demonstrated, it remains unclear why different masking strategies lead to varying performance gains, particularly in terms of semantic differences in the information retained or reintroduced by each strategy. To investigate this, we analyze it from two perspectives: information diversity and information reliability. Information diversity is quantified using shannon entropy, which reflects the amount of uncertainty or richness in the image content(Hjelm et al. 2018). In addition, the reliability of the information is assessed by deep feature similarity, indicating how semantically aligned the reused image is with the original image(Zhang et al. 2018). This analysis reveals how the reuse branch contributes to effective feature learning under different masking strategies. Eq.3 defines the shannon entropy of the image I. For RGB inputs, we compute it on the grayscale intensity histogram over $[0, 255]$.

$$H(I) = -\sum_{i=0}^{255} p(i) \log_2 p(i) \tag{3}$$

where $p(i)$ denotes the normalized frequency of the pixel value i in the image. A higher entropy value indicates greater information diversity. The difference in entropy between the masked image and the reused image (ΔH) reflects the amount of new information introduced by the reuse branch. Table 8 shows the differences in the information entropy between different masked images and reused images.

As shown in Table 8, the Patch and Grid methods lead to larger entropy gains, suggesting that they introduce more diverse content into the learning process. In contrast, the Random method results in a smaller entropy difference, indicating limited additional information, which may explain its relatively lower effectiveness.

Method	H_m	H_c	ΔH
Patch	6.224	7.147	0.923
Grid	6.307	7.243	0.936
Random	6.297	6.911	0.614

Table 8: Comparison of Information Entropy of Different Masking Methods

Method	Model	S_{ds}	$\mathcal{F}$
Patch		0.537	0.943
Grid	ResNet-101	0.706	0.875
Random		0.426	0.771

Table 9: Deep feature cosine similarity and final $\mathcal{F}$ scores obtained using Different Masking methods

However, entropy alone does not capture the semantic quality of the reused information. For example, a large difference in information entropy may still be observed even if the reused image undergoes a significant domain shift, but the information may lack reliability. To complement this, we evaluate the deep feature similarity. Specifically, we use a pre-trained ResNet-101 model to extract high-level features from both the original and reused images, and compute the cosine similarity. This result is transformed into a similarity score using Eq.4, with the corresponding results summarized in Table 9.

$$S\left(I, I_c\right) = e^{-\left|S_{\text{ds}}(I, I_c) - S_{\text{a}}\right|} \tag{4}$$

where S_{ds} represents the deep feature similarity between the reused image and the original image based on deep features, while S_{a} is a hyperparameter that defines the optimal similarity value for deep features, which is set to 0.5 by default. To quantify this trade-off, we define a evaluation metric that combines entropy difference and similarity, defined as Eq.5

$$F\left(I, I_m, I_c\right) = \frac{1}{N} \sum_{i=1}^{N} w_1 S\left(I, I_c\right) + w_2 \Delta H\left(I_m, I_c\right) \tag{5}$$

where N represents the number of sample pairs, which is set to 1000 in this case. I, I_1^m and I_c^i represent the original image, the masked image, and the reuse image, respectively. w_1 and w_2 represent different weight values, both set to 0.5 by default. The final results are presented in Table 9.

As shown in Table 9, the Grid method achieves the highest similarity due to its fixed sampling pattern, but introduces structural redundancy that reduces feature sensitiViTy. Conversely, the Random method yields low similarity, causing feature conflicts that impede learning. The Patch method attains moderate similarity through optimal diversity-reliability trade-offs, achieving superior performance (highest $\mathcal{F}$).

Feature Visualization

To intuitively understand the effectiveness of the proposed method, we visualize the intermediate feature representations of the baseline model (ResNet-34) and our model. As shown in Figure 5, ResNet-34 tends to lose fine-grained structural details in the shallow layers (e.g. the chimney of the ship), causing blurred boundaries and imprecise localization in deeper layers. In contrast, MaskResNet-34 adopts a dual-branch architecture with an information reuse mechanism, which strengthens the model's sensitiViTy to local structures and enables better preservation of object boundaries and textures in the early stages. In deeper layers, MaskAnyNet-34 produces more semantically aligned and spatially precise activations, indicating improved understanding of both object location and shape.

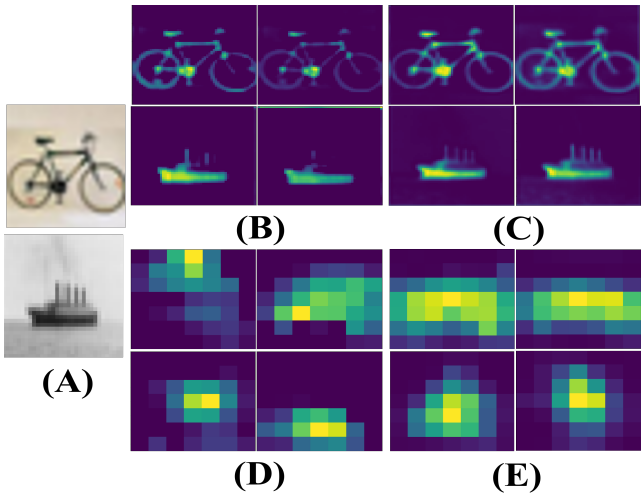


Figure 5: Feature visualization comparisons of ResNet-34 and our MaskResNet-34. (A) Original image, (B) Low-level features from ResNet-34, (C) Low-level features from MaskResNet-34, (D) High-level features from ResNet-34, (E) High-level features from MaskResNet-34.

These observations confirm that our method not only enables baseline models to better retain local details but also promotes a more robust global semantic understanding of scene context and category relationships, thereby enriching feature representations throughout the model hierarchy.

Conclusions

In this paper, we propose MaskAnyNet, a general visual architecture that addresses the limitations of traditional masking strategies in supervised learning, which commonly discard masked regions as noise, forfeiting fine-grained details. The proposed MaskAnyNet introduces a reuse branch to relearn masked content and fuse it with the masked features, enhancing structural detail and feature diversity. This dual-branch design captures complementary global semantics and local intricacies, thereby significantly improving the generalization capabilities of the model. Extensive experiments on multiple benchmarks demonstrate consistent performance gains over state-of-the-art baselines. Further analysis highlights the importance of masking and fusion strategies, providing insight for future visual representation learning.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62373324, U24A20242, 62475241 and 62271448), in part by the Zhejiang Province Leading Geese Plan (Grant No. 2025C02160), in part by the National Key Research and Development Program of China (Grant No. 2024YFC3306902).

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, C.-F. R.; Fan, Q.; and Panda, R. 2021. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 357–366.

Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; and Chan, S.-H. G. 2023. Run, don't walk: chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12021–12031.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chen, P.; Liu, S.; Zhao, H.; Wang, X.; and Jia, J. 2020. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*.

Chen, X.; Ding, M.; Wang, X.; Xin, Y.; Mo, S.; Wang, Y.; Han, S.; Luo, P.; Zeng, G.; and Wang, J. 2024. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision*, 132(1): 208–223.

Chi, C.; Zhang, S.; Xing, J.; Lei, Z.; Li, S. Z.; and Zou, X. 2019. Selective refinement network for high performance face detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8231–8238.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeVries, T. 2017. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv preprint arXiv:1708.04552*.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; Yu, N.; and Guo, B. 2023. Peco: Perceptual codebook for bert pre-training of vision transformers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 552–560.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.

Feng, Z.; and Zhang, S. 2024. Evolved Hierarchical Masking for Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ghiasi, G.; Lin, T.-Y.; and Le, Q. V. 2018. Dropblock: A regularization method for convolutional networks. *Advances in neural information processing systems*, 31.

Girdhar, R.; El-Nouby, A.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10406–10417.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Jocher, G.; Qiu, J.; and Chaurasia, A. 2023. Ultralytics YOLO.

Kumar Singh, K.; and Jae Lee, Y. 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE international conference on computer vision*, 3524–3533.

Li, P.; Li, X.; and Long, X. 2020. Fencemask: a data augmentation approach for pre-extracted image features. *arXiv preprint arXiv:2006.07877*.

Li, S.; Zhang, L.; Wang, Z.; Tian, J.; Tan, C.; Liu, Z.; Yu, C.; Xie, Q.; Lu, H.; Wang, H.; et al. 2025. Mergevq: A unified framework for visual generation and representation with disentangled token merging and quantization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19713–19723.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, Y.; Zhang, S.; Chen, J.; Chen, K.; and Lin, D. 2023. Pixmim: Rethinking pixel reconstruction in masked image modeling. *arXiv preprint arXiv:2303.02416*.

Liu, Z.; Hao, Z.; Han, K.; Tang, Y.; and Wang, Y. 2024. GhostNetV3: Exploring the Training Strategies for Compact Models. *arXiv preprint arXiv:2404.11202*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Murdock, C.; Li, Z.; Zhou, H.; and Duerig, T. 2016. Blockout: Dynamic model selection for hierarchical deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2583–2591.

Peng, Z.; Dong, L.; Bao, H.; Ye, Q.; and Wei, F. 2022. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*.

Qin, C.-J.; Wu, R.-Q.; Liu, Z.; Lin, X.; Guo, C.-L.; Park, H. H.; and Li, C. 2024. Restore anything with masks: Leveraging mask image modeling for blind all-in-one image restoration. In *European Conference on Computer Vision*, 364–380. Springer.

Qin, D.; Leichner, C.; Delakis, M.; Fornoni, M.; Luo, S.; Yang, F.; Wang, W.; Banbury, C.; Ye, C.; Akin, B.; et al. 2025. MobileNetV4: Universal Models for the Mobile Ecosystem. In *European Conference on Computer Vision*, 78–96. Springer.

Shikhar, S.; and Sobti, A. 2024. Label-free Anomaly Detection in Aerial Agricultural Images with Masked Image Modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5440–5449.

Snoek, C. G.; Worring, M.; and Smeulders, A. W. 2005. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, 399–402.

Tan, M.; and Le, Q. 2021. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106. PMLR.

Vishniakov, K.; Xing, E.; and Shen, Z. 2022. MixMask: Revisiting Masking Strategy for Siamese ConvNets. *arXiv preprint arXiv:2210.11456*.

Wang, X.; Nie, G.; Meng, J.; and Yan, Z. 2025. MIMTrack: In-Context Tracking via Masked Image Modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7979–7987.

Wang, X.; Shrivastava, A.; and Gupta, A. 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2606–2615.

Wei, C.; Fan, H.; Xie, S.; Wu, C.-Y.; Yuille, A.; and 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*, 14668–14678.

Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; and Yan, S. 2017. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1568–1576.

Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I. S.; and Xie, S. 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16133–16142.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Neural Information Processing Systems (NeurIPS)*.

Xie, Z.; Zhang, Z.; Cao, Y.; Lin, Y.; Bao, J.; Yao, Z.; Dai, Q.; and Hu, H. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9653–9663.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Zhan, Y.; Zhao, Y.; Luo, C.; Zhang, Y.; and Sun, X. 2023. Attention-Guided Contrastive Masked Image Modeling for Transformer-Based Self-Supervised Learning. In *2023 IEEE International Conference on Image Processing (ICIP)*, 2490–2494.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Ma, Y.; Fan, F.; Zhang, Y.; and Huang, J. 2017. Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. *Journal of the Optical Society of America A*, 34(8): 1400–1410.

Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; and Chen, J. 2024. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16965–16974.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13001–13008.

Zhou, B.; Zhao, H.; Puig, X.; Xiao, T.; Fidler, S.; Barriuso, A.; and Torralba, A. 2019. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127: 302–321.

Zhou, Z.; Wang, B.; Li, S.; and Dong, M. 2016. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters. *Information fusion*, 30: 15–26.