ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240493

MIM-HD: Making Smaller Masked Autoencoder Better with Efficient Distillation

Zherui Zhang[1], Changwei Wang[2,4], Rongtao Xu[3], Wenhao Xu[1], Shibiao Xu [1,*], Li Guo[1], Jiguang Zhang[3], Xiaoqiang Teng[1] and Wenbo Xu[1]

[1]Artificial Intelligence, Beijing University of Posts and Telecommunications
[2]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences)
[3]Institute of Automation,Chinese Academy of Sciences
[4]Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science

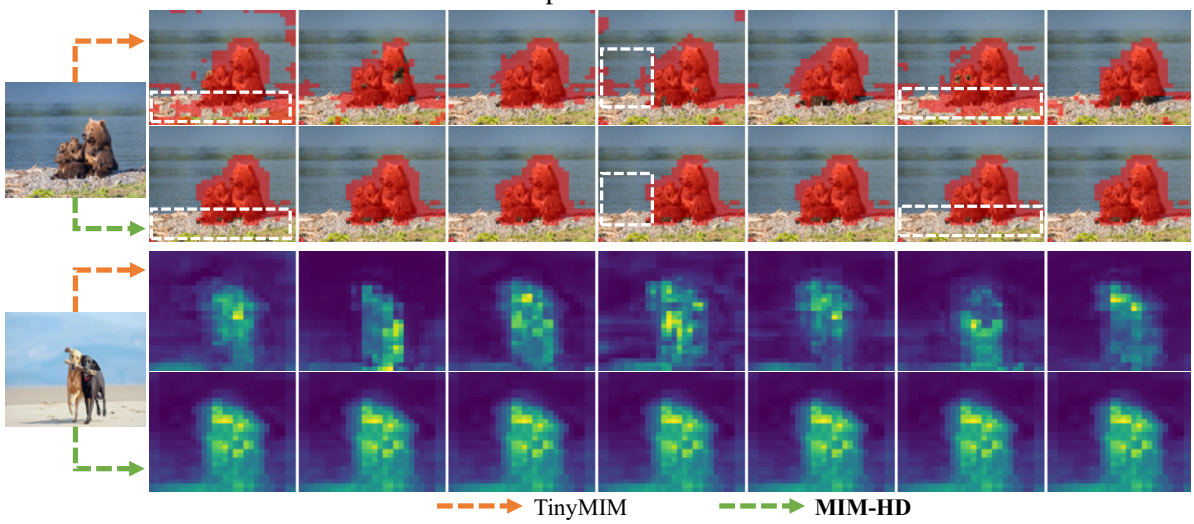

Figure 1: Attentive properties within the head. Inspired by DINO (https://github.com/facebookresearch/dino), the visualization approach reveals that each head of our MIM-HD concentrates more on semantically discriminative regions compared to the TinyMIM, and effectively eliminates background noise, resulting in a more powerful student.

Abstract. Self-supervised learning and knowledge distillation intersect to achieve exceptional performance on downstream tasks across diverse network capacities. This paper introduces MIM-HD, which implements enhancements for masked image modeling (MIM) distillation, in two key aspects. First, a vision transformer head-level relation adaptive distillation approach is proposed, allowing the student to dynamically draw multi-source knowledge from the teacher based on its evolving state, compatible with scenarios where teacher-student transformer block head count differs. Second, to address the overemphasis on the encoder and neglect of the decoder role in maintaining representation consistency in previous MIM distillations, a dual-view decoding strategy for latent visual representations is introduced, reusing the teacher's decoder to alleviate MIM burdens on smaller networks. MIM-HD effectiveness is demonstrated through evaluations on ADE20K (mIoU) and ImageNet-1K (Acc), achieving +1.4% and +0.5% improved performance, respectively, compared to state-of-the-art methods, with substantial advantages on smaller pre-training datasets. Moreover, MIM-HD achieves superior efficiency, reducing pre-training epochs from 300 to 100.

* Corresponding Author: shibiaoxu@bupt.edu.cn

1 Introduction

Profound intrinsic representations can be learned in a self-supervised manner from unlabeled visual datasets, allowing models to develop general understanding without reliance on explicit task labels. Masked image modeling (MIM), an established self-supervised pre-training approach, has demonstrated its efficacy through outstanding performance when fine-tuned on downstream tasks spanning image classification [28, 31], semantic segmentation[33, 32, 29, 30, 22, 24], and multimodal understanding[34, 38]. Masked autoencoders (MAE)[8] exemplify this family of techniques, utilizing an encoder-decoder architecture to reconstruct corrupted regions of input images.

However, accessing these promising properties is not without cost. Firstly, MIM does not alleviate the computational demands of pre-training large vision transformer(ViT)[7] from scratch using extensive datasets and complex optimization procedures. Secondly, larger model capacities seem to derive disproportionately greater advantage from MIM compared to smaller architectures, with MIM potentially hindering performance in more compact models. This may

stem from MIM serving as a challenging reconstruction task that overtaxes smaller models' abilities to establish meaningful semantic associations between visible and masked regions, thereby impeding coherent decoding and recovery of raw pixels.

The TinyMIM[18] study explores enabling smaller models to benefit from MIM pre-training by systematically investigating various distillation factors when using MIM-trained teachers, adopting token relations as the distillation target. Specifically for self-attention derived relations, a mismatch exists in the common multi-head setting, for example, 12 heads in MAE-Base vs 16 heads in MAE-Large. TinyMIM proposes replacing an adaptive block in the student encoder tail layer to align teacher-student head counts, then performing head-to-head relation alignment as shown in Figure 2(c). However, this approach still has the following three limitations:

1. **Limited Knowledge Source.** The student receives knowledge for each head only from a single pre-determined teacher head, lacking exposure to a teacher's full representational breadth.
2. **Neglect of Student Evolution.** This approach neglects the student's developmental process by potentially overburdening limited early-stage imitation abilities.
3. **Optimization Complexity.** The introduction of the adaptive block inconsistently alters optimization complexity between model components, with potential negative effects, especially when extending it beyond the tail layer.

In summary, we argue that the efficient distillation process should allow students to dynamically acquire knowledge from teachers based on its own developmental states, without disrupting model optimization properties. Beyond TinyMIM, related MAE distillation work (see Sec. 2) excessively focuses on the encoder while neglecting the decoder's role in maintaining representation consistency, decoupling from the objectives of MIM pretraining.

In this paper, we introduce MIM-HD, a novel and more efficient distillation framework that further facilitates the utilization of MIM by smaller models. Our main contributions are as follows :

- **Adaptability and Compatibility.** we propose a transformer block head-level adaptive distillation approach that enables the student to obtain attentive representations from multi-head of the teacher in an adaptive manner. Such a design not only makes the distillation process student-driven, but also resolves potential issues arising from mismatched ViT block head counts.
- **Visual representation consistency.** To alleviate the complexity of the MIM pretext task for smaller models, we propose a dual-view decoding strategy. This strategy performs the generation of complementary regions on the output token feature of the student's encoder and reuses the pre-trained decoder of the larger MAE for raw pixel reconstruction and maintaining visual representation consistency.
- **Significant improvement.** our extensive experiments conducted on ADE20K and ImageNet-1K datasets validate the effectiveness of MIM-HD. For ViT-B, our approach improves the mIoU metric by **+2.6** and **+0.4** compared to MAE and TinyMIM, respectively. Furthermore, our approach exhibits potential benefits in the domain of vision transformer token pruning by providing more precise hints of important regions, as shown in Figure 1.
- **Efficient training.** Evaluation experiments demonstrate that MIM-HD achieves more efficient knowledge transfer in resource-constrained scenarios, requiring only 100 pre-training epochs.

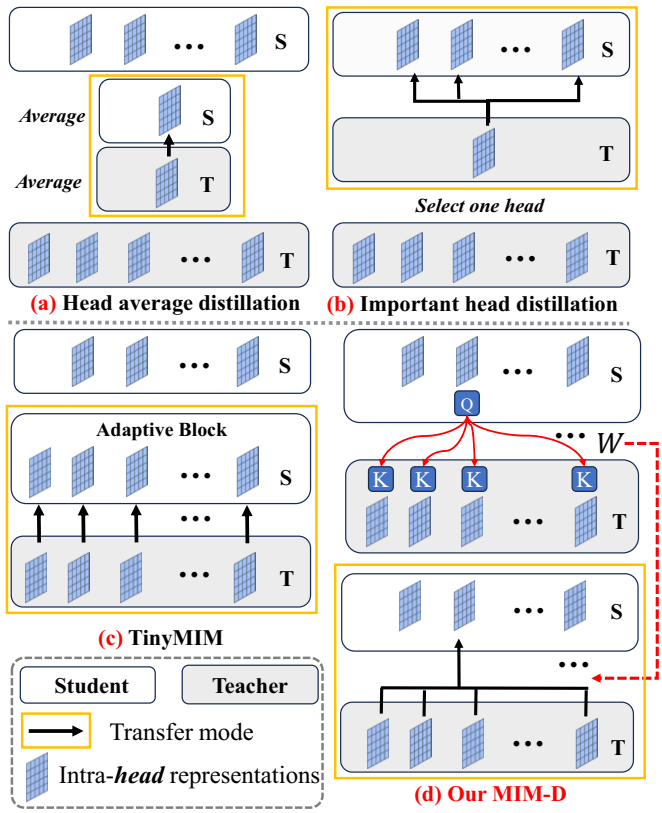

Figure 2: Common solutions to address head number mismatch: (a) perform averaging operations on multi-head of teachers and students, respectively; (b) select the head that exerts the most significant influence on the downstream task for the teacher; (c) utilize the adaptive block introduced by TinyMIM to align the head number; (d) we introduce MIM-HD, which effectively extracts knowledge from the multi-head of the teacher by generating adaptive weights, ensuring compatibility even in scenarios with mismatched head counts.

2 Related Work

Within the knowledge distillation paradigm[37], models with stronger performance are considered teachers for training more compact student networks, enabling the preservation of outstanding capabilities while streamlining architectural requirements.

MAE distillation. At the intersection realm of self-supervision and knowledge distillation, there is an absence of task-specific knowledge transfer, such as logit dark knowledge in classification tasks, thus necessitating a greater emphasis on extracting more valuable intermediate visual representations from the encoder and subsequently achieving efficient transfer. Masked Image Modeling (MIM) is a powerful self-supervised pre-training method, and Masked Autoencoders (MAE)[8] are a representative visual solution within this domain. Approaches such as DMAE[1], G2SD[9], and TinyMIM[18] implement direct knowledge transfer from the teacher encoder, incorporating techniques like token feature alignment and token relation transfer. TinyMIM empirically examines the diverse factors affecting distillation while decoupling the decoder from MAE. G2SD categorizes knowledge into task-agnostic and task-specific, subsequently performing two-stage distillation. In AMD[39] and MGD[36] student networks, the decoder employs the output of the teacher's encoder as the reconstruction target. Additionally, methods such as SMKD[13], MixMAE[15], and SdAE[6] integrate the MIM with mask-level Mixup online to construct a teacher, thereby mitigat-

ing spatial redundancy while exploiting latent visual representations, however, these methods are still applied to the image pixel space. Unlike native MAE, SD-MAE[16] offers supplementary supervision to the visible patches.

Inadequately explored. Despite the substantial advancements in performance facilitated by these methods, several areas remain underexplored: ***i)*** prior methods have underutilized or neglected transferring insights from the teacher's decoder, despite its potential contribution to the student's ability to reconstruct inputs. ***ii)*** inadequate exploration of distillation at the head-level for transformer blocks composed of multi-head. In the SSTA[26] for ***ii)***, the head that exerts the greatest influence on accuracy is explored as the primary source of knowledge. However, this approach is constrained by the specific downstream task and, within the context of a self-supervised paradigm, ignores the discriminative representations embedded within the other heads. Our MIM-HD can benefit from more than one head. Furthermore, we advocate for the reuse of the decoder of the teacher network to mitigate the dilemma encountered in TinyMIM.

3 Method

3.1 Adaptive and Compatible Relation Distillation

Inspired by the research conducted on TinyMIM[18], which calculates token relations and acts as targets for distillation through the pair form of **V-V** and **Q-K** derived from the transformer block, makes it possible for smaller networks to reap the self-supervised benefits of the pre-trained larger MIM networks. For the m_{th} head of the transformer block at layer-i, the above token relation pairs can be written as $R^{VV}_{i,m}$,$R^{QK}_{i,m}$, which can be formulated as:

$$\begin{aligned} R^{VV}_{i,m} &= Softmax(\frac{V^m_i V^{m\top}_i}{\sqrt{D/M}}), \\ R^{QK}_{i,m} &= Softmax(\frac{Q^m_i K^{m\top}_i}{\sqrt{D/M}}), \end{aligned} \tag{1}$$

where $R^{VV}_{i,m}, R^{QK}_{i,m} \in \mathbb{R}^{N\times N}$, D and M denote the dimension of the latent space and the head number, respectively, and the [class] token is ignored for simplicity.

It is crucial to acknowledge that these relation pairs are independently calculated within the respective networks of teachers and students, which necessitates addressing the challenge of the mismatch in head number(M). In addition to the TinyMIM solution, we report in Section 1 and Figure 2(c). In Figure 2, we summarize several common solutions. Specifically, Figure 2(a) represents the averaging strategy, which involves performing head-level distillation by merging multi-head into a single head, and in AttnDistill[25] involves transfer [class] token attention weights via the similar averaging strategy, yet this coarse fusion ignores the diversity of head representations. Figure 2(b) depicts the selection of the most significant head in SSTA[26] by quantifying its impact on accuracy. However, SSTA is intertwined with a specific downstream task.

To mitigate the aforementioned issue, we propose a student-driven head-level adaptive distillation approach. Specifically, drawing inspiration from self-attention, we generate adaptive weights by treating the student's multi-head and the teacher's multi-head as independent queries and keys, respectively, as illustrated in the upper part of Figure 2(d). Subsequently, these weights are utilized for the knowledge transfer of token relations. This setup is not only compatible with the scenario of mismatch in the head number between teachers and students, but also allows students to dynamically draw knowledge from multiple sources.

Figure 3 provides an overview of the proposed method. To formulate our adaptive matching mechanism in head-level distillation, we consider distilling the V-V relation pair from the i_{th} transformer block of teacher to the j_{th} transformer block of student and, for more general purposes, denote the head number within the block as M_1(teacher head number) and M_2(student head number), respectively, which means:

$$R^{VV(T)}_i \rightarrow R^{VV(S)}_j, \tag{2}$$

where $R^{VV(T)}_i \in \mathbb{R}^{M_1\times N\times N}$, $R^{VV(S)}_j \in \mathbb{R}^{M_2\times N\times N}$ and $M_1 \neq M_2$, thus the candidate set of heads is generated:

$$\mathbb{C} = \{(m_1, m_2)|\forall\, m_1 \in [1, \cdots, M_1], m_2 \in [1, \cdots, M_2]\}, \tag{3}$$

for instance, $R^{VV(T)}_{i,m_1}$ denotes V-V relation pair from the m_1-th head of the i_{th} transformer block of the teacher, while $R^{VV(S)}_{j,m_2}$ denotes the m_2-th head of the j_{th} transformer block of the student. For the sake of clarity in notation, K-K and Q-V relation pairs all follow similar conventions, let us uniformly denote the relation distillation as $R^{(T)} \rightarrow R^{(S)}$, after selecting the corresponding matching layer for teachers and students. In scenarios involving teacher-student layer mismatch, we adopt the initial setup from TinyMIM, where the target is designated as the 18_{th} block of the teacher encoder.

Inspired by self-attention mechanism, our head-level adaptive distillation approach considers the multi-head of student as queries, dynamically assigning attention weights to the multi-head of teacher, as shown in Figure 2(d). This enables the student to selectively acquire specific knowledge based on its current state of evolution. Notably, pairwise similarity matrices can be employed to quantify the semantic similarity among the heads, as demonstrated in SemCKD[23]:

$$\begin{aligned} S^{(T)} &= F(R^{(T)}) \cdot F(R^{(T)})^\top, \\ S^{(S)} &= F(R^{(S)}) \cdot F(R^{(S)})^\top, \end{aligned} \tag{4}$$

where $F(\cdot) : \mathbb{R}^{M\times N\times N} \longmapsto \mathbb{R}^{M\times 1}$ represents the average operation, while $S^{(T)} \in \mathbb{R}^{M_1\times M_1}$ and $S^{(S)} \in \mathbb{R}^{M_2\times M_2}$ denote the similarity matrices of teacher and student, respectively.

For the student, m_2-th head, the query vector of the subspace $q^S_{m_2}$ is obtained by feeding the similarity matrix $S^{(S)}$ into the projection module $Proj$:

$$q^{(S)}_{m_2} = Proj(S^{(S)}), \tag{5}$$

where $q^{(S)}_{m_2} \in \mathbb{R}^{1\times C}$, C is the embedding dimension of the query subspace, similar operations are applied to the other student heads to obtain the final query vector $q^{(S)} \in \mathbb{R}^{M_2\times C}$. Following the same process, we can obtain the key vector $k^{(T)} \in \mathbb{R}^{M_1\times C}$ from the teacher similarity matrix $S^{(T)}$. Finally, we can obtain the head-level attention weight matrix between students and teachers:

$$\mathbf{W}_{(m2,m1)} = \frac{e^{q^{(S)}_{m2}\cdot k^{(T)\top}_{m1}}}{\sum_j e^{q^{(S)}_{m2}\cdot k^{(T)\top}_j}}, \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{M_2\times M_1}$, and the generated attention weights, and satisfy condition $\sum_{m_1=1}^{M_1} \mathbf{W}_{(m2,m1)} = 1$, $(m1, m2) \in \mathbb{C}$. Once prepared, the objective function of the distillation relation pair $\mathcal{L}_R$ is

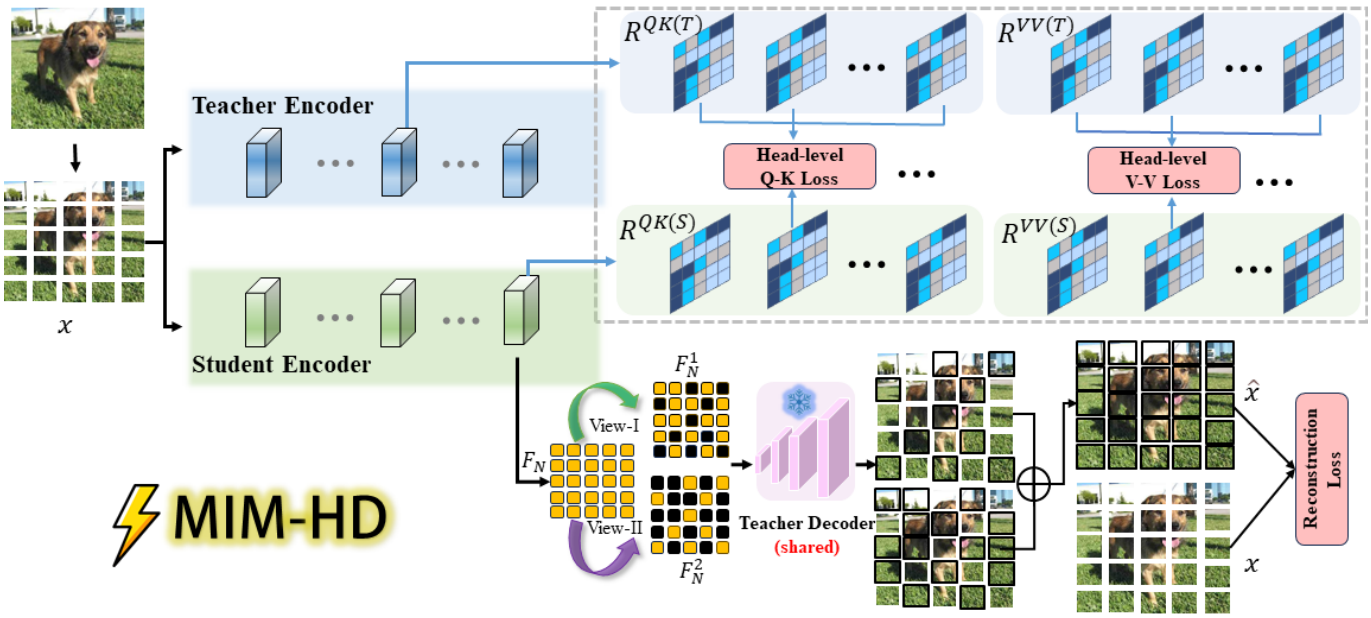

Figure 3: Overview of the proposed MIM-HD. The left half of the figure illustrates the proposed head-level adaptive distillation approach employed during the transfer of token relation pair knowledge. This approach enables students to dynamically acquire multi-head knowledge and is compatible with scenarios involving a mismatch in the head number. The right half depicts our proposed dual-view decoding strategy, which enhances the consistency of visual representations by reusing the pre-trained teacher decoder. This strategy allows smaller models to benefit from the advantages of MIM.

obtained:

$$\begin{aligned}\mathcal{L}_R &= \sum_{(m_1,m_2)\in\mathbb{C}} Dist(R^{(S)}_{m_2}, R^{(T)}_{m_1}, \mathbf{W}_{(m2,m1)}) \\ &= \sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2} Dist(R^{(S)}_{m_2}, R^{(T)}_{m_1}, \mathbf{W}_{(m2,m1)}) \\ &= \sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2} KL(R^{(S)}_{m_2}, R^{(T)}_{m_1}, \mathbf{W}_{(m2,m1)}),\end{aligned} \tag{7}$$

where KL is Kullback-Leibler divergence. Specifically, we denote the relation pair objective functions for Q-K and V-V as $\mathcal{L}_R^{QK}$ and $\mathcal{L}_R^{VV}$, respectively.

3.2 *Benefit from MIM*

In previous research on MAE distillation, some approaches[1, 9, 18] enable students to fully replicate the visual representations from a single perspective, potentially over-emphasizing the encoder at the expense of the decoder. Alternatively, directly applying the masked autoencoding pre-training paradigm to smaller networks does not consistently provide the expected gains in downstream task performance and may even prove inferior to training from scratch.

Upon reviewing these confusions, we argue that precisely mimicking a teacher's modeling capabilities requires students to emulate both the teacher's encoder and decoder. To achieve this, we design an approach wherein the output of the students' encoder is masked and provided as input to the shared pre-trained teacher decoder, which then performs reconstruction of the raw pixel data as it does during the teacher's own pre-training. Compared to the MAE training paradigm, which generates discrete regions at the encoder input and performs masking there, the visual representations generated by the visible regions are independent of the masked ones, introducing inconsistent mask symbols during fine-tuning. The weak correlations between the two types of regions make the reconstruction of these mask regions challenging for the smaller network. It is important to note that in our approach, the entire image serves as input to the student encoder, resulting in semantically strongly relevant visual representations. Furthermore, the similarity of the tokens in the deeper layers of the encoder is also highly remarkable[35, 21, 19], indicating that the visual representations captured within the tokens are no longer independent of each other. Consequently, the visible tokens possess the ability to generate raw pixels in the neighboring masked region, and our configuration potentially alleviates the difficulty associated with executing the MIM task on smaller networks.

To further leverage the powerful recovery capability of the pre-trained teacher decoder and enhance the consistency of visual representations, we propose a simple yet effective complementary dual-view recovery mechanism. Specifically, we randomly generate a mask $\mathcal{V}_1$ with a drop rate of 50% as one of the views and designate the remaining region as mask $\mathcal{V}_2$, the other view. Subsequently, applied to the student encoder output $F_N \in \mathbb{R}^{N\times D}$, the [class] token is omitted here:

$$\begin{cases} F_N^1 = [F_N \odot \mathcal{V}_1, F_m], \\ F_N^2 = [F_N \odot \mathcal{V}_2, F_m] \end{cases} \tag{8}$$

where symbol [ ] denotes the concatenation operation, $F_N^1 \in \mathbb{R}^{N\times D}$ and $F_N^2 \in \mathbb{R}^{N\times D}$ represent the visual representations of the tokens generated under the dual-view mask, respectively, and F_m are the shared learnable mask tokens. To achieve raw pixel recovery, we take both F_N^1 and F_N^2 to the shared pre-trained teacher decoder $\mathbf{Dec}$, respectively, following the design guidelines of the MAE, we also only calculate the reconstruction loss in the mask region:

$$\hat{X} = \mathcal{R}(\mathbf{Dec}(F_N^2)\odot\mathcal{V}_1, \mathbf{Dec}(F_N^1)\odot\mathcal{V}_2), \tag{9}$$

$$\mathcal{L}_{REC} = \|X - \hat{X}\|_2^2, \tag{10}$$

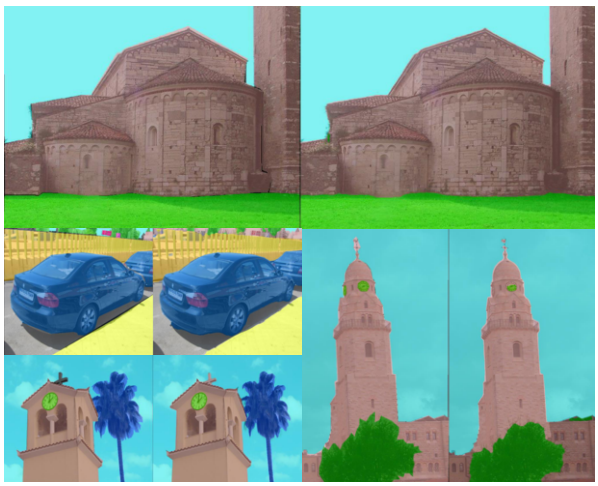

Figure 4: Visualization of ADE20K segmentation results. The left figure represents the ground truth, while the right figure represents the output of MIM-HD.

where $\mathcal{R}(\cdot)$ is unshuffle operation used to restore the positions of the visible regions and masked regions, $\hat{X}$ represents the recovered raw pixels, and $\odot$ denotes element-wise multiplication. Finally, the reconstruction loss $\mathcal{L}_{REC}$ is computed using the Mean Squared Error(MSE).

3.3 Optimization

To leverage the advantages of the MIM pre-training paradigm during knowledge distillation for smaller networks, MIM-HD introduces specific behaviors outlined in Eq. 7 and Eq. 10. These equations embody head-level token relation-based distillation in the encoder stage and dual-view raw pixel reconstruction in the decoder stage, respectively. The combination of these strategies leads to the final joint optimization training objective, denoted as $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_R^{QK} + \mathcal{L}_R^{VV} + \alpha\mathcal{L}_{REC}, \tag{11}$$

where α is the regularization factor.

4 Experiments

4.1 Implementation Details

In the experimental section, we adopt an end-to-end strategy to optimize the parameters of the student network, guided by the objective function defined in Eq. 11. During the pre-training phase, we utilize the AdamW optimizer with an initial learning rate of $1.5e^{-3}$ and employ a cosine annealing strategy for the learning rate policy. The image input size is set to 224×224. For the fine-tuning downstream task phase, our experiments are conducted on NVIDIA RTX 3090 GPUs and the OpenMMLab platform. Table 2 presents a summary of the statistics for the downstream tasks that we employed for validation purposes.

4.2 Downstream Tasks

We evaluate the visual representations learned from MAE-ViT-L using the proposed MIM-HD on the two downstream tasks presented in Table 2. For specific task heads used, we employ a linear layer for classification and follow UperNet[27] for semantic segmentation, respectively.

As can be observed from Table 1, our MIM-HD demonstrates superior performance compared to TinyMIM across various capacities of the student model, even with a pre-training phase of only 100 epochs. For instance, in the ImageNet classification(Top-1 Acc), utilizing our approach for knowledge transfer on ViT-B outperforms MAE, CAE, and TinyMIM by **+1.2**, **+0.9**, and **+0.3**, respectively. Similarly, in the ADE20K semantic segmentation(mIoU), our approach outperforms MAE, CAE, and TinyMIM by **+2.6**, **+0.5**, and **+0.4**, respectively, and we show some segmentation results in Figure 4.

Furthermore, during the pre-training phase, we explore the potential of MIM-HD to facilitate knowledge extraction on constrained datasets, specifically working with TinyImageNet. As can be seen from the gray-marked area, our method surpasses TinyMIM by **+1.4** mIoU on ADE20K.

4.3 Visualization

The discriminative properties of the transformer block multi-head in TinyMIM and MIM-HD are visualized by analyzing the attention weights of the [class] token in the last layer, following the validation scheme demonstrated in DINO[3]. The characterization patterns of the first 8 heads are presented in Figure 5, marked using orange and green arrows for TinyMIM and MIM-HD, respectively. The following observations can be obtained:

1. **Incomplete semantic regions**: In TinyMIM (orange arrow), some transformer block heads capture only partial semantic regions of the object or background noise, as observed in the dashed box of Figure 5.
2. **Stronger head behavior**: In our MIM-HD (green arrow), each transformer block head independently draws knowledge from multi-head of the teacher model. This distributed approach facilitates the identification of semantically meaningful regions, even for complex objects or scenes, as each student transformer block head can focus on a valuable aspect of the object concept or domain.

The enhanced representational capabilities of MIM-HD directly benefit **token pruning** techniques[12, 17, 2, 10] that rely on analyzing [class] token attention weights. More precisely, by providing more accurate region markings, MIM-HD facilitates the selective retention of informative image tokens while helping to mitigate the inclusion of noisy or irrelevant image tokens.

4.4 Ablation Study

To validate the effectiveness of each component of the proposed method, we start by pre-training ViT-B on the TinyImageNet dataset and then conduct various ablation experiments on ADE20K.

Ablation studies of our MIM-HD framework: To isolate the influence of various factors on distillation performance, we adopt the TinyMIM initial configuration, employing a ViT-B student encoder with its last layer mimicking the MAE-ViT-L teacher encoder on the 18_{th} layer. In Table 3, we establish the distillation targets as the Q-K and V-V relation pairs. Subsequently, by applying our head-level adaptive distillation approach with $\mathcal{L}_R^{QK}$ and $\mathcal{L}_R^{VV}$ as objective functions, we outperform TinyMIM by 1.2 on ADE20K. We experiment with replacing the distance function with MSE, but this does not result in any performance gains. Finally, the implementation of the $\mathcal{L}_{REC}$ objective function corresponding to the dual-view strategy

Table 1: Comparison with state-of-the-art MIM methods. Accuracy and mean Intersection over Union (mIoU) are used as evaluation metrics in the downstream ImageNet-1K classification task and ADE20K segmentation task, respectively. ^{T}: model pre-training on TinyImageNet. *: the reproduction results on 2 NVIDIA 3090 RTX GPUs using the officially released code.

Backbone	Method	Pre-training dataset	Pre-training epochs	Segmentation mIoU(%)	Classification Top-1 Acc(%)
ViT-B	MAE[8]$_{CVPR'22}$	IN1K	1600	48.1	83.6
	CAE[5]$_{IJCV'22}$	DALLE250M+IN22K+IN1K	1600	50.2	83.9
	SdAE[6]$_{ECCV'22}$	IN1K	300	48.6	84.1
	DINO[3]$_{ICCV'21}$	IN1K	1600	47.2	83.3
	Ge^2-AE[14]$_{AAAI'23}$	IN1K	800	48.9	84.8
	A^{2}MIM[11]$_{ICML'23}$	IN1K	800	49.0	84.2
	TinyMIM*[18]$_{CVPR'23}$	IN1K	300	50.3	84.5
	MIM-HD(Ours)	IN1K	**100**	**50.7**(+0.4)	**84.8**(+0.3)
	TinyMIM^{*T}[18]$_{CVPR'23}$	TinyImageNet	300	39.6	79.2
	MIM-HD^{T}(Ours)	TinyImageNet	**100**	**41.0**(+1.4)	**79.7**(+0.5)
ViT-S	MAE[8]$_{CVPR'22}$	IN1K	1600	42.8	80.6
	DINO[3]$_{ICCV'21}$	IN1K	1600	45.3	81.5
	TinyMIM*[18]$_{CVPR'23}$	IN1K	300	46.8	82.3
	MIM-HD(Ours)	IN1K	**100**	**47.6**(+0.8)	82.3
ViT-Ti	MAE[8]$_{CVPR'22}$	IN1K	1600	37.6	71.6
	Moco[4]$_{ICCV'21}$	IN1K	1600	39.3	73.3
	TinyMIM*[18]$_{CVPR'23}$	IN1K	300	42.8	75.2
	MIM-HD(Ours)	IN1K	**100**	**43.2**(+0.4)	**75.7**(+0.5)

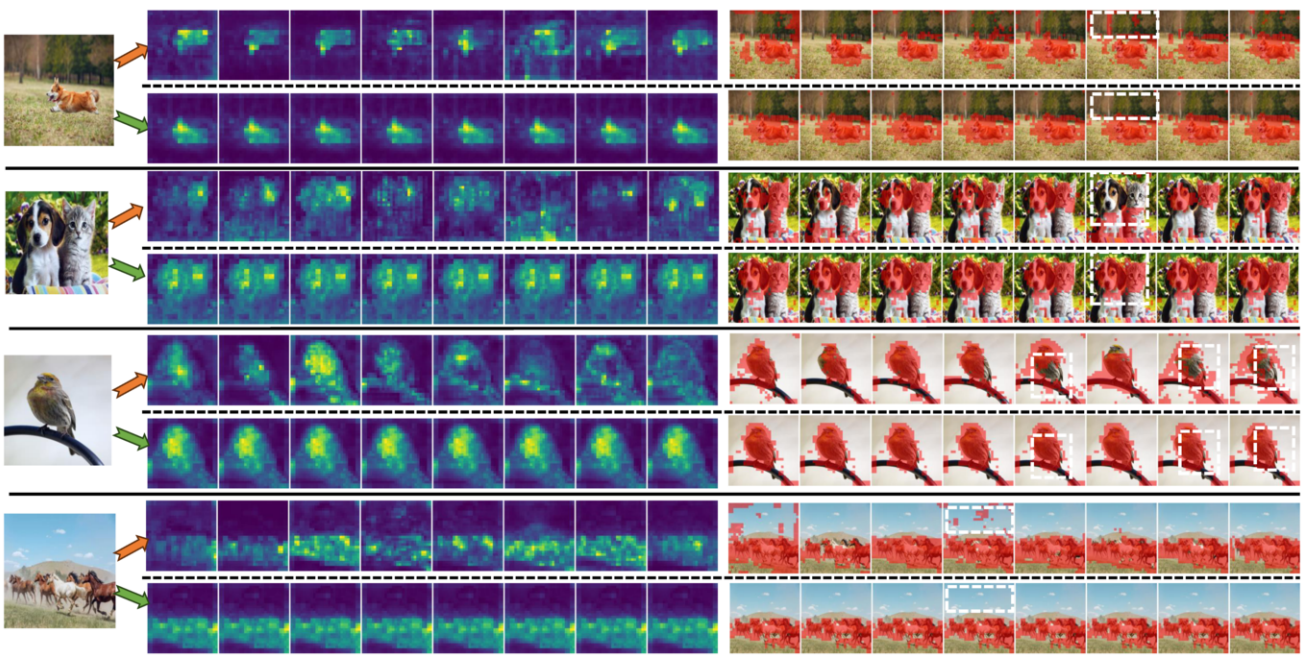

Figure 5: Visualization of the attention area maps within the student ViT-B last layer transformer block first 8 heads and the selection of semantic regions. The orange arrow and The green arrow represent TinyMIM and MIM-HD (Ours), respectively. In the dashed box, it can be observed that TinyMIM suffers from issues such as semantic region incompleteness and background noise. Our approach, allocates significant attention to the discriminative regions in all the heads.

Table 2: Dataset Statistics

Type	Dataset	#Classes
Classification	ImageNet-1K [20]	1,000
Semantic segmentation	ADE20K [40]	150

yields a further improvement in mIoU, $40.8 \rightarrow 41.0$, thus validating the effectiveness of our proposed method.

Ablation studies on teacher target block: In Table 4, we report the impact of varying target depths in the pre-training phase on the downstream task. When the stacking depths of the teacher and student networks differ, we conduct experiments with the S-9_{th} and S-12_{th} layers of the student network taking the T-10_{th}, T-12_{th}, T-18_{th}, and T-24_{th} layers of the teacher network as distillation targets, respectively. An empirical setup is proposed in TinyMIM and DMAE[1], specifi-

Table 3: Ablation experiments with our MIM-HD components.

$\mathcal{L}_R^{QK}$	$\mathcal{L}_R^{VV}$	$\mathcal{L}_{REC}$	KL	MSE	Method	*mIoU*
					TinyMIM[18]$_{CVPR'23}$	39.6
✔			✔			40.3
✔				✔	**MIM-HD**	40.1
✔	✔				**(Ours)**	40.8
✔	✔	✔				**41.0**

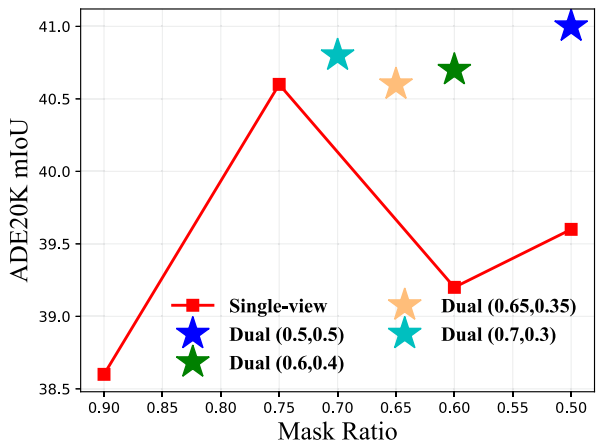

Figure 6: Ablation experiments with dual-view decoding strategies. At different mask ratios, dual-view decoding consistently outperforms single-view decoding for dense prediction tasks, such as ADE20K.

cally, T-18$_{th}$ → S-12$_{th}$ in TinyMIM, as well as in DMAE where it is discovered to be more effective at $\frac{3}{4}$ depth, that is, T-18$_{th}$ → S-9$_{th}$. It can be observed that our MIM-HD yields superior performance when T-18$_{th}$ → S-12$_{th}$.

Table 4: Ablation experiments with teacher target block

Student layer	Teacher layer			
	10_{th}	12_{th}	$18_{th}(\frac{3}{4})$	24_{th}
$9_{th}(\frac{3}{4})$	37.8	39.2	40.7	40.1
12_{th}	38.1	39.6	**41.0**	40.3

Ablation studies on dual-view strategy: Our dual-view decoding strategy aims to further enhance the consistency of the teacher-student encoder output by reusing the generative power of the shared pre-trained teacher decoder to recover the raw pixel. To substantiate the effectiveness of this method, we conduct comparative analyses under varying mask ratios for both single-view and dual-view configurations, as depicted in Figure 6. Notably, our dual-view approach exhibits superior performance in downstream tasks. This can be primarily attributed to two factors: firstly, the reduced complexity for smaller models in executing the MIM task, as discussed in Section 3.2, and secondly, the extraction of region-complementary visual representations from multiple perspectives, which proves advantageous for downstream intensive prediction tasks.

5 Further analysis

5.1 *Training speed of MIM-HD*

Table 5 details the pre-training time consumed by the proposed MIM-HD model on ImageNet-1K. The metrics reported are: ***i)*** **PT.Ep** - the number of iterative epochs required for pre-training; ***ii)*** **EP.Hours** - the average time consumed per epoch, calculated over five independent runs; ***iii)*** **Total.Hours** - the full training time of the entire network is calculated cumulatively.

Table 5: To elaborate on the efficiency comparison between TinyMIM and MIM-HD (Ours), the training time for a single epoch (Ep. Hours) is obtained by averaging the results of five independent executions to account for variability, while the total time (Total Hours) is calculated cumulatively.

Method	Model	PT.Ep	Ep. Hours	Total.Hours	Device
TinyMIM[18]$_{CVPR'23}$	ViT-B	300	2.4	720	3090
MIM-HD (**Ours**)	ViT-B	**100**(-67%)	2.8	**280**(-61%)	3090
TinyMIM[18]$_{CVPR'23}$	ViT-S	300	1.8	540	3090
MIM-HD (**Ours**)	ViT-S	**100**(-67%)	2.2	**220**(-59%)	3090

While our approach introduces a slight increase in computational overhead per epoch, stemming primarily from computing the Q-K and V-V weight distribution matrices W and lightweight decoder inference, these additional costs are outweighed by more efficient knowledge extraction and transfer, thus making it easier for the student network to converge. Specifically, compared to TinyMIM, our MIM-HD requires fewer pre-training epoch, reducing the number of iterations from 300 → 100. Consequently, the overall training duration is reduced by approximately 60%, achieving comparable downstream task performance with TinyMIM.

5.2 *Future work*

This paper aims to achieve efficient distillation by enhancing the method of knowledge extraction and transfer from the teacher network. Figure 1 and Figure 5 illustrate that our approach yields more precise importance region indications, suggesting its potential use as a weight initialization method that directly benefits ViT token pruning and token merging domains[12, 17, 2, 10].

6 Conclusion

In this paper, we propose the **MIM-HD** distillation framework, which enables smaller models to further capitalize on the pre-training advantages of masked image modeling. It comprises two primary components: an **adaptive head-level token relation distillation** approach and a **dual-view decoding** strategy. The former allows student models to acquire intermediate visual representations from more than one transformer block head of the teacher based on its evolutionary state, and compatible with scenarios where the number of student and teacher multi-head differ. The latter involves performing a dual-view mask on the encoder output to mitigate the difficulty faced by smaller models in executing the masked image modeling task by reusing the shared pre-trained teacher decoder. The effectiveness of our approach is validated through experiments on various downstream tasks. Moreover, our method is significantly more relaxed with respect to pre-training configurations than state-of-the-art masked image modeling distillation methods, reducing training time by around **60%** while achieving competitive downstream task performance.

Acknowledgements

This work is supported by Beijing Natural Science Foundation No. JQ23014, in part by the National Natural Science Foundation of China (Nos. 62271074, 62071157, 62302052, 62171321 and 62162044)

References

[1] Y. Bai, Z. Wang, J. Xiao, C. Wei, H. Wang, A. L. Yuille, Y. Zhou, and C. Xie. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24256–24265, 2023.

[2] D. Bolya, C.-Y. Fu, X. Dai, P. Zhang, C. Feichtenhofer, and J. Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

[3] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

[4] X. Chen*, S. Xie*, and K. He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021.

[5] X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022.

[6] Y. Chen, Y. Liu, D. Jiang, X. Zhang, W. Dai, H. Xiong, and Q. Tian. Sdae: Self-distillated masked autoencoder. In *European Conference on Computer Vision*, pages 108–124. Springer, 2022.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[9] W. Huang, Z. Peng, L. Dong, F. Wei, J. Jiao, and Q. Ye. Generic-to-specific distillation of masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15996–16005, 2023.

[10] Z. Kong, P. Dong, X. Ma, X. Meng, W. Niu, M. Sun, X. Shen, G. Yuan, B. Ren, H. Tang, et al. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *European Conference on Computer Vision*, pages 620–640. Springer, 2022.

[11] S. Li, D. Wu, F. Wu, Z. Zang, and S. Z. Li. Architecture-agnostic masked image modeling – from vit back to cnn. In *International Conference on Machine Learning*, 2023.

[12] Y. Liang, C. Ge, Z. Tong, Y. Song, J. Wang, and P. Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=BjyvwnXXVn_.

[13] H. Lin, G. Han, J. Ma, S. Huang, X. Lin, and S.-F. Chang. Supervised masked knowledge distillation for few-shot transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19649–19659, 2023.

[14] H. Liu, X. Jiang, X. Li, A. Guo, Y. Hu, D. Jiang, and B. Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1649–1656, 2023.

[15] J. Liu, X. Huang, J. Zheng, Y. Liu, and H. Li. Mixmae: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6252–6261, 2023.

[16] Y. Luo, Z. Chen, S. Zhou, and X. Gao. Self-distillation augmented masked autoencoders for histopathological image classification. *arXiv preprint arXiv:2203.16983*, 2022.

[17] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

[18] S. Ren, F. Wei, Z. Zhang, and H. Hu. Tinymim: An empirical study of distilling mim pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3687–3697, 2023.

[19] L. Ru, H. Zheng, Y. Zhan, and B. Du. Token contrast for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2023.

[20] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[21] Y. Tang, K. Han, Y. Wang, C. Xu, J. Guo, C. Xu, and D. Tao. Patch slimming for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12165–12174, 2022.

[22] C. Wang, R. Xu, Y. Zhang, S. Xu, and X. Zhang. Retinal vessel segmentation via context guide attention net with joint hard sample mining strategy. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1319–1323. IEEE, 2021.

[23] C. Wang, D. Chen, J.-P. Mei, Y. Zhang, Y. Feng, and C. Chen. Semckd: semantic calibration for cross-layer knowledge distillation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[24] C. Wang, R. Xu, S. Xu, W. Meng, and X. Zhang. Da-net: Dual branch transformer and adaptive strip upsampling for retinal vessels segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 528–538. Springer, 2022.

[25] K. Wang, F. Yang, and J. van de Weijer. Attention distillation: self-supervised vision transformer students need more guidance. *arXiv preprint arXiv:2210.00944*, 2022.

[26] H. Wu, Y. Gao, Y. Zhang, S. Lin, Y. Xie, X. Sun, and K. Li. Self-supervised models are good teaching assistants for vision transformers. In *International Conference on Machine Learning*, pages 24031–24042. PMLR, 2022.

[27] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer, 2018.

[28] Z. Xie, Z. Zhang, Y. Cao, Y. Lin, J. Bao, Z. Yao, Q. Dai, and H. Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022.

[29] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang. Dc-net: Dual context network for 2d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 503–513. Springer, 2021.

[30] R. Xu, Y. Li, C. Wang, S. Xu, W. Meng, and X. Zhang. Instance segmentation of biological images using graph convolutional network. *Engineering Applications of Artificial Intelligence*, 110:104739, 2022.

[31] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, and X. Zhang. Self correspondence distillation for end-to-end weakly-supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3045–3053, 2023.

[32] R. Xu, C. Wang, S. Xu, W. Meng, and X. Zhang. Wave-like class activation map with representation fusion for weakly-supervised semantic segmentation. *IEEE Transactions on Multimedia*, 2023.

[33] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang. Rssformer: Foreground saliency enhancement for remote sensing land-cover segmentation. *IEEE Transactions on Image Processing*, 32:1052–1064, 2023.

[34] R. Xu, J. Zhang, J. Sun, C. Wang, Y. Wu, S. Xu, W. Meng, and X. Zhang. Mrftrans: Multimodal representation fusion transformer for monocular 3d semantic scene completion. *Information Fusion*, page 102493, 2024.

[35] Y. Xu, Z. Zhang, M. Zhang, K. Sheng, K. Li, W. Dong, L. Zhang, C. Xu, and X. Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2964–2972, 2022.

[36] Z. Yang, Z. Li, M. Shao, D. Shi, Z. Yuan, and C. Yuan. Masked generative distillation. In *European Conference on Computer Vision*, pages 53–69. Springer, 2022.

[37] D. Zhang, H. Li, W. Cong, R. Xu, J. Dong, and X. Chen. Task relation distillation and prototypical pseudo label for incremental named entity recognition. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 3319–3329, 2023.

[38] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and W. He. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.

[39] Z. Zhao, B. Huang, S. Xing, G. Wu, Y. Qiao, and L. Wang. Asymmetric masked distillation for pre-training small foundation models. *arXiv preprint arXiv:2311.03149*, 2023.

[40] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.