Generic-to-Specific Distillation of Masked Autoencoders

Wei Huang[1,*], Zhiliang Peng[1,§,*], Li Dong[2], Furu Wei[2], Jianbin Jiao[1,†], Qixiang Ye[1,†]

University of Chinese Academy of Sciences[1]

Microsoft Research[2]

Abstract

Large vision Transformers (ViTs) driven by self-supervised pre-training mechanisms achieved unprecedented progress. Lightweight ViT models limited by the model capacity, however, benefit little from those pre-training mechanisms. Knowledge distillation defines a paradigm to transfer representations from large (teacher) models to small (student) ones. However, the conventional single-stage distillation easily gets stuck on task-specific transfer, failing to retain the task-agnostic knowledge crucial for model generalization. In this study, we propose generic-to-specific distillation (G2SD), to tap the potential of small ViT models under the supervision of large models pre-trained by masked autoencoders. In generic distillation, decoder of the small model is encouraged to align feature predictions with hidden representations of the large model, so that task-agnostic knowledge can be transferred. In specific distillation, predictions of the small model are constrained to be consistent with those of the large model, to transfer task-specific features which guarantee task performance. With G2SD, the vanilla ViT-Small model respectively achieves 98.7%, 98.1% and 99.3% the performance of its teacher (ViT-Base) for image classification, object detection, and semantic segmentation, setting a solid baseline for two-stage vision distillation. Code will be available at https://github.com/pengzhiliang/G2SD.

1. Introduction

Vision transformers (ViTs) [11, 55] have been promising representation models, particularly when trained upon large-scale datasets using self-supervised learning methods [5]. The masked image modeling (MIM) methods [4, 13], which train representation models by reconstructing pixels [13, 52, 57], tokens [4, 8, 35] or features [2, 47], promoted the performance of large ViT models to a new height.

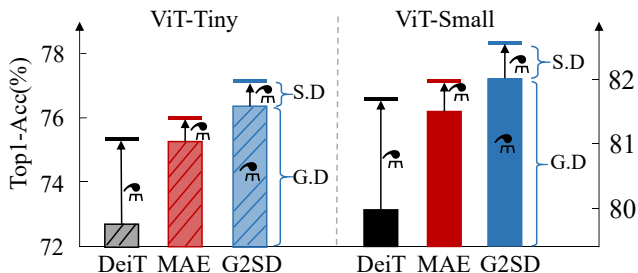However, when acclaiming the promising performance of large ViT models, we notice that small ViT models,

Figure 1. Comparison of single-stage distillation models (from scratch [41] and pre-trained by the self-supervised method MAE [13]) with the two-stage distillation counterparts (G2SD) using the same teacher model. G.D and S.D respectively denote generic and specific distillation. 🐎 is the symbol of distillation.

e.g., ViT-Tiny and ViT-Small, unfortunately, benefit little from either the big training data or self-supervised learning methods. For example, the ViT-Large model trained by MAE [13] outperforms the CNN model [29] by 1.6 points on ImageNet-1k, while the ViT-Small model is inferior to its CNN counterpart [29]. In most scenarios with limited computational resources, *e.g.*, front-end recognition systems, CNNs [15, 19] remain the preferred models.

Do vanilla small ViT models really have no future? We attempt to answer this question from the perspective of knowledge distillation in this study. To fulfill this purpose, the first step is to revisit the conventional knowledge distillation methods [18, 38, 41] in the age of supervised learning. It is observed that task-oriented distillation [41] reports unsatisfactory performance, Fig. 1. One reason could be that this kind of task-oriented distillation only focus on task-specific knowledge while missing some kind of task-agnostic knowledge which is beneficial to generalization ability improvement and can be effectively endowed by self-supervised teacher model. In natural language processing, two-stage distillation method, *e.g.*, Tiny-BERT [22], was exploited to overcome the limitation and transfer generic knowledge embedded from teacher to student models. Nevertheless, whether or not this paradigm applicable to vision tasks remains unexplored.

In this study, we aim to establish a general-to-specific

---

*Equal contribution. § Contribution during internship at Microsoft Research. † Corresponding authors.

distillation baseline for vision tasks based on sophisticated self-supervised learning (*e.g.*, MAE [13]), to guarantee that lightweight ViTs can simultaneously soak up task-agnostic and task-specific representations from teacher models for greater generalization and higher task performance, Fig. 1. Specifically, at the generic distillation stage, a student model is encouraged to obtain the task-agnostic knowledge from the teacher models. The encoder and decoder of pre-trained MAE constitute the teacher model while a light-weight decoder is attached to the lightweight vision Transformer as the student model, Fig. 2. The input image is randomly partitioned to visible and masked patches. The visible patches are fed to encoders. The hidden feature outputs of teacher decoder's intermediate layer is used to guide training of the student model. For task-specific distillation, the fine-tuned MAE model equipped with task layers [14, 41, 51] teaches student model the task-specific knowledge (*e.g.*, classification score). The student backbone is initialized from the previous distillation stage while the task layers are randomly initialized. Predictions of the student are constrained to be consistent with those of the teacher as well as ground truth labels. Such a task-specific distillation phase guarantees the performance of downstream tasks, *e.g.*, image classification, object detection and semantic segmentation.

With G2SD, the vanilla ViT-Small model with **26%** parameters and **2.6×** throughput of the ViT-Base teacher, obtains 1) **98.6%** (82.5% *vs*. 83.6%) top-1 accuracy of its teacher on ImageNet-1k [39] for image classification task, 2) **98.1%** (50.6 *vs*. 51.6) mAP of its teacher on MS COCO [26] for object detection and 3) **99.3%** (48.0 *vs*. 48.3) mIoU of its teacher on ADE20k [58] for semantic segmentation. Furthermore, G2SD demonstrates better generalization ability than its single-stage distillation counterparts in terms of occlusion invariance and robustness.

The contributions are summarized as follows:

- We propose general-to-specific distillation (G2SD) to transfer task-agnostic and task-specific knowledge from masked autoencoders to lightweight ViTs, setting a solid baseline for two-stage vision model distillation.

- We design a simple-yet-effective generic distillation strategy by aligning the student's predictions with hidden features of the pre-trained masked autoencoder at visible and masked patches.

- Experiments show that the lightweight student model with G2SD achieves competitive results across vision tasks, improving the performance of lightweight ViT models to a new height.

2. Related Work

Vision Transformers. ViTs [11] have achieved impressive performance across vision tasks [4,13,25,35,36,55,56]. Furthermore, ViTs demonstrated the superiority in terms of robustness and generalization [13, 32, 35], compared to their CNN counterparts. However, due to the lack of inductive bias, ViTs report unsatisfactory performance in the limited model capacity regime [11, 41]. One solution is to explicitly introduce convolutional operators to ViTs [31,50] to enhance the competitiveness compared to lightweight CNNs [19]. The other way is using large models act as teachers to transfer inductive bias to ViTs in the knowledge distillation fashion [7, 41, 50]. This study focuses on the latter.

Self-supervised Learning. To explore big data without high-quality labels, self-supervised learning has been the preferred paradigm to construct representation models [5]. Masked language modeling [10] achieved great success in natural language process (NLP) field. Inspired by it, BEiT [4] introduced the *mask-then-predict* paradigm to the computer vision filed and exploited the great potential of masked image modeling (MIM) on various tasks. BEiT v2 [35] constructed a semantic-rich visual tokenizer in order to get better target. MAE [13] set a new baseline for MIM by reconstructing pixels at mask patches with a decoupling encoder-decoder architecture. Meanwhile, feature masking and reconstruction methods [2,47] demonstrated advantages over the pixel-reconstruction approach. Those methods, however, when exploring performance upper bound by finding better supervisions to pre-train large ViTs, ignored the adaptability of lightweight models with limited capacity. In this study, G2SD develops a two-stage knowledge distillation baseline for lightweight ViT models to enjoy MIM advantages.

Knowledge Distillation. The pioneering work [18] compressed the "dark knowledge" from a large (teacher) model to a small (student) model by minimizing KL divergence between the output logits distribution of the two models. NKD [53] rethinked the relation between knowledge distillation loss and the original cross-entropy and proposed a new KD loss. FitNet [38] pioneered feature distillation by utilizing the intermediate layers' features from the teacher model. To find the better feature layers for distillation, subsequent works [6, 20] studied the factor of connection path cross multiple i ntermediate layers between teacher and student networks. Besides distilling the knowledge contained in samples, inter-samples relation, as structural information, was transferred to student models [33, 34, 43]. Knowledge distillation has also been elaborately studied for ViTs [7, 21, 27, 41, 54]. SSTA [49] simultaneously learned
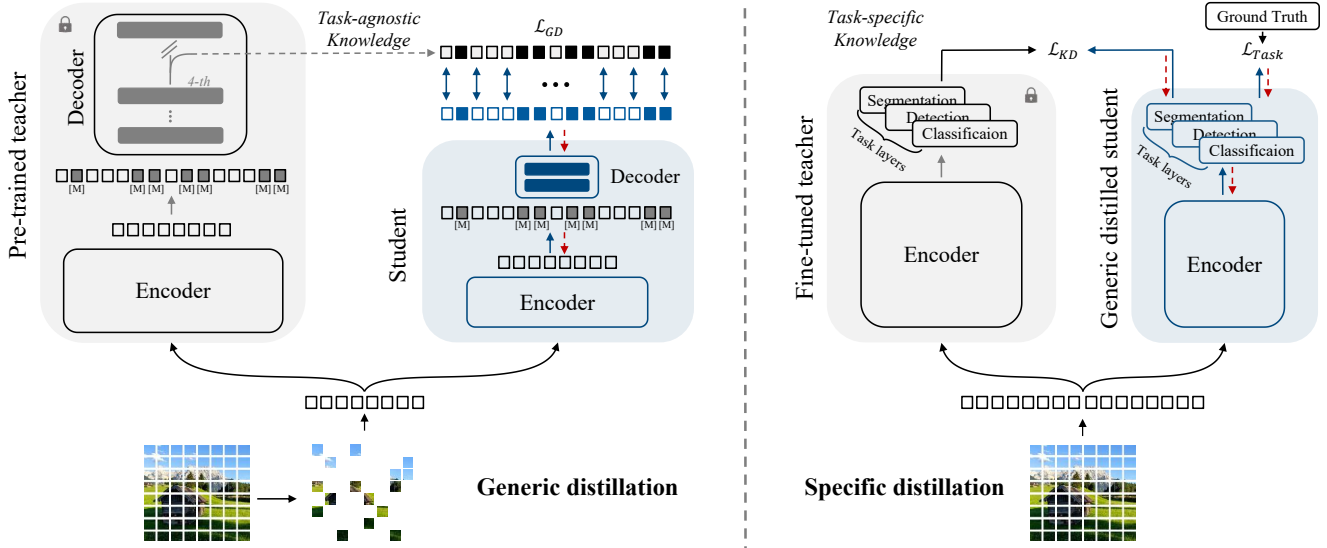
Figure 2. Diagram of the proposed generic-to-specific distillation (G2SD). [M] denotes mask token. In the generic distillation stage (*left*), masked images are converted to patches and fed to both the teacher and student encoders for feature extraction. Feature predictions of the student decoder are aligned with those of the teacher at both visible and predicted patches. In the specific distillation stage (*right*), student models are trained to have consistent predictions with teacher models fine-tuned on the specific task.

from the supervised teacher and self-supervised teacher, which was regarded as teaching assistant. However, those methods are designed and evaluated on specific task, such as classification, detection [12] or segmentation [40]. The task-oriented methods experience difficulty in transferring task-agnostic knowledge, while task-agnostic knowledge is crucial to guarantee the generalization ability of lightweight models. To overcome the limitation, TinyBERT [22] pioneered two-stage knowledge distillation in natural language processing. Nevertheless, the problem remains to be explored in vision tasks. In this study, we focus on excavating task-agnostic knowledge embedded in masked autoencoders to establish a solid baseline for vision model distillation in the era of self-supervised learning.

3. Preliminary

Transformer Representations. To learn visual representations, ViT [11] converts each image to a sequence of 'words' (vectors) by partitioning it to patch grid. In specific, the input image $x \in \mathbb{R}^{H \times W \times C}$ is divided to $N = (H * W)/P^2$ non-overlapping patches $\{x_i^p\}_{i=1}^N$, where H, W, C and P respectively denote the image height, width, channel and patch stride, and $x_i^p \in \mathbb{R}^{N \times (P^2 C)}$. In this study, a $224 \times 224 \times 3$ size image is reshaped to a 14×14 grid of image patches, each patch size is $16 \times 16 \times 3$. Meanwhile, positional information[1] is embedded to the patches. By passing the vectors through stacked Transformer blocks, which consist of a multi-head self-attention [44] layer and a

fully connected feed-forward network, the input vectors are converted to image representations.

Masked Autoencoders. The MAE model contains an encoder f_e and a decoder f_d, where both f_e and f_d are stacked Transformer blocks. The input tokens $\{x_i^p\}_{i=1}^N$ are grouped to the visible token set $\{x_i^p\}_{i \in \mathcal{V}}$ and the masked token set $\{x_i^p\}_{i \in \mathcal{M}}$. While the visible tokens are fed to the encoder f_e to extract features, the masked tokens act as the learning targets, which are required to be reconstructed during self-supervised learning (MIM). In the MAE method [13], a high mask ratio (*e.g.*, 75%) is adopted, to prevent information leakage (*i.e.*, simply extrapolating masked pixels from the neighbors) in the pre-training phase.

Specifically, $\{x_i^p\}_{i \in \mathcal{V}}$ are fed to f_e to obtain latent features $\{e_i\}_{i \in \mathcal{V}}$, where $e_i = f_e(x_i^p)$ for each $i \in \mathcal{V}$. A shared learnable mask token $e_{[M]}$ is considered as the placeholder of tokens in $\mathcal{M}$. After that, we have the input tokens $\{h_i\}_{i=1}^N$ for the decoder f_d, where

$$h_i = e_{[M]} \odot \delta(i \in \mathcal{M}) + e_i \odot (1 - \delta(i \in \mathcal{M})), \quad (1)$$

and $\delta(\cdot)$ denotes an indicator function. $\{h_i\}_{i=1}^N$ are then fed to f_d to generate predictions at all positions $\{z_i\}_{i=1}^N$. The loss is calculated by comparing the normalized pixels with predictions at masked positions $\mathcal{M}$, as

$$\mathcal{L}_{\text{MAE}} = \sum_{i \in \mathcal{M}} ||\text{LN}(x_i^p) - z_i||_2, \quad (2)$$

where $\text{LN}(\cdot)$ is the layer normalization without affine transformation, *a.k.a*, the per-patch normalization in MAE. Af-

---
[1] The positional embeddings are omitted for simplicity

ter pre-training, the encoder acts as backbone to extract representations for various tasks and the decoder is abandoned. As the model does not access any label in the pre-training stage, it is assumed that the features extracted by encoder are general to downstream tasks.

4. Generic-to-Specific Distillation

Our generic-to-specific distillation (G2SD) emphasizes transferring the task-agnostic knowledge embedded in large pre-trained masked autoencoders [13]. In conjunction with task-specific distillation, G2SD endows lightweight models favorable generalization ability and competitive results.

4.1. Generic Distillation: Task-agnostic knowledge Transfer

In each training iteration, the generic distillation consists of a feed-forward procedure of the teacher model, a feed-forward and a back-propagation procedure of the student model, Fig. 2 (*left*). In the feed-forward procedure, outputs from an intermediate layer of the teacher decoder and the final layer of the student decoder are compared to calculate the generic distillation loss.

Denote the encoder and decoder of teacher model pre-trained with MAE method as f_e^t and f_d^t, and the encoder and decoder of student models as f_e^s and f_d^s, respectively. Input tokens $\{\boldsymbol{x}_i^p\}_{i=1}^N$ are randomly categorized to visible ones $\{\boldsymbol{x}_i^p\}_{i\in\mathcal{V}}$ and masked ones $\{\boldsymbol{x}_i^p\}_{i\in\mathcal{M}}$. The visible tokens $\{\boldsymbol{x}_i^p\}_{i\in\mathcal{V}}$ are simultaneously fed to f_e^t and f_e^s to extract features $\{\boldsymbol{e}_i^t\}_{i\in\mathcal{V}}$ and $\{\boldsymbol{e}_i^s\}_{i\in\mathcal{V}}$. According to Eq. (1), we have the input tokens set $\{\boldsymbol{h}_i^t\}_{i=1}^N$ for the teacher decoder and $\{\boldsymbol{h}_i^s\}_{i=1}^N$ for the student decoder. In general, a flexible decoder consists of multiple Transformer blocks. We respectively mark the depth of teacher decoder and that of student decoder as L and l, where $l \leq L$ in our experiments. Let features output by the l-th layer of the teacher decoder f_d^t as $\{\hat{\boldsymbol{z}}_i^t\}_{i=1}^N$, where $\hat{\boldsymbol{z}}_i^t = f_{d_l}^t(\boldsymbol{h}_i^t)$. The student decoder f_d^s employs l Transformer blocks on $\{\boldsymbol{h}_i^s\}_{i=1}^N$ and calculates the output features as $\{f_d^s(\boldsymbol{h}_i^s)\}_{i=1}^N$. Subsequently, a linear layer $\boldsymbol{W}$ is applied on $f_d^s(\boldsymbol{h}_i^s)$ to align with the channel dimension of $\hat{\boldsymbol{z}}_i^t$ and generates predictions $\boldsymbol{z}_i^s$, *i.e.*, $\boldsymbol{z}_i^s = \boldsymbol{W}f_d^s(\boldsymbol{h}_i^s)$.

According to the above definitions, a generic distillation loss is defined as

$$\mathcal{L}_{\text{GD}} = \sum_{i\in\{\mathcal{V}\bigcup\mathcal{M}\}} \text{Smooth-}\ell_1(\text{LN}(\hat{\boldsymbol{z}}_i^t) - \boldsymbol{z}_i^s), \quad (3)$$

where Smooth-$\ell_1(\cdot)$ is a trade-off function between ℓ_1 and ℓ_2. By minimizing $\mathcal{L}_{\text{GD}}$ on the visible tokens $\mathcal{V}$, the student encoder is optimized to extract features in the way like the teacher encoder, *i.e.*, mimicking feature extraction behavior. By minimizing $\mathcal{L}_{\text{GD}}$ on the masked tokens $\mathcal{M}$, the student encoder and decoder are optimized to learn context

modeling ability from teacher models. Optimizing $\mathcal{L}_{\text{GD}}$ on all tokens transfers task-agnostic knowledge.

4.2. Specific Distillation: Task-specific Representation Configuration

After generic distillation, lightweight models are able to generalize to downstream tasks and reach competitive performance, which has been validated by comprehensive experiments (See Tab. 4). Nevertheless, limited by a relatively small model size and number of parameters, lightweight models still have a performance gap with their teachers. To bridge the gap, specific distillation is performed so that compact yet discriminative features can be configured for downstream tasks, such as image classification, object detection, and semantic segmentation.

For specific distillation, the teacher model f^t is first pre-trained with MAE method then fine-tuned on the specific task. A lightweight ViT model after generic distillation is set as the student f^s. As concrete the loss function is depend on specific tasks, we denote $\mathcal{L}_{\text{Task}}$ as the task loss function, $\mathcal{L}_{\text{KD}}$ as the task-specific distillation loss function. Combining the task loss with task-specific distillation loss, we have a joint loss to optimize the student model, as

$$\mathcal{L}_{\text{SD}} = \mathcal{L}_{\text{Task}}(f^s(\boldsymbol{x}), Y) + \beta\mathcal{L}_{\text{KD}}(f^s(\boldsymbol{x}), f^t(\boldsymbol{x})), \quad (4)$$

where Y is the ground truth and β is the regularization factor (Refer to Appendix A for details).

4.3. Analysis

The proposed two-stage approach is more plausible than commonly used single-stage methods, which can be justified from the perspective of mutual information [1]. The knowledge distillation can be generally interpreted as a procedure to maximize the mutual information $\mathcal{I}$ of a teacher model (f^t) and a student model (f^s). Denote the parameters of the student model as θ^s, the pre-training dataset as X and the fine-tuning dataset as $\hat{X}$. The single-stage task-specified distillation is interpreted as

$$\arg\max_{\theta^s}\mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|\hat{X}), \quad (5)$$

which maximizes the mutual information between the teacher model f^t and the student model f^s conditional on the fine-tuning dataset $\hat{X}$. The proposed G2SD is interpreted as

$$\arg\max_{\theta^s}\mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|X) + \mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|\hat{X})$$
$$- \mathcal{I}_{\theta^s,\theta^t}(f^t, f^s|(X, \hat{X})), \quad (6)$$

which maximizes the mutual information between the teacher model f^t and the student model f^s conditional on both the pre-training data X and the fine-tuning dataset $\hat{X}$.

Table 1. Top-1 accuracy on ImageNet-1k.

Method	Teacher	#Param(M)	Acc (%)
DeiT-Ti [41]		5	72.2
MobileNet-v3 [19]		5	75.2
ResNet-18 [15]		12	69.8
DeiT-S [41]		22	79.8
BEiT-S [4]	N/A	22	81.7
CAE-S [8]		22	82.0
DINO-S [5]		22	82.0
iBOT-S [59]		22	82.3
ResNet-50 [15]		25	76.2
Swin-T [28]		28	81.3
ConvNeXt-T [29]		29	82.1
DeiT-Ti⚗ [41]		6	74.5
DeiT-S⚗ [41]	RegNetY-	22	81.2
DearKD-Ti [7]	16GF	6	74.8
DearKD-S [7]		22	81.5
Manifold-Ti [21]		6	75.1
Manifold-S [21]	CaiT-	22	81.5
MKD-Ti [27]	S24	6	76.4
MKD-S [27]		22	82.1
SSTA-Ti [49]	DeiT-S	6	75.2
SSTA-S [49]	DeiT-B	22	81.4
DMAE-Ti [3]		6	70.0
DMAE-S [3]	MAE-B	22	79.3
G2SD-Ti (ours)		6	77.0
G2SD-S (ours)		22	**82.5**

Table 2. Object detection and instance segmentation results on the MS COCO dataset.

Method	#Param(M)	AP^{bbox}	AP^{mask}
Mask R-CNN [14], 36 epochs + Multi-Scale			
CAE-S [8]	46.1	44.1	39.2
ViT-Adapter-T [9]	28.1	46.0	41.0
Swin-T [28]	47.8	46.0	41.6
ConvNeXt-T [29]	48.1	46.2	41.7
imTED-S [56]	30.1	48.0	42.8
ViT-Adapter-S [9]	47.8	48.2	42.8
ViTDet [25], 100 epochs + Single-Scale			
DeiT-S⚗ [41]	44.5	47.2	41.9
DINO-S [5]	44.5	49.1	43.3
iBOT-S [59]	44.5	49.7	44.0
G2SD-Ti (ours)	27.7	46.3	41.6
G2SD-S (ours)	44.5	**50.6**	**44.8**

Table 3. ADE20K validation results using UperNet [51]. The input image resolution is 512×512.

Method	#Param(M)	mIoU
ViT-Adapter-Ti [9]	36.1	42.6
Swin-T [28]	59.9	44.5
ConvNeXt-T [29]	60	46.0
ViT-Adapter-S [9]	57.6	46.6
DINO-S [5]	42.0	44.0
iBOT-S [59]	42.0	45.4
G2SD-Ti (ours)	11.0	44.5
G2SD-S (ours)	42.0	**48.0**

Obviously, the mutual information defined by Eq. (6) is larger than that by Eq. (5), which implies more information can be transferred by our G2SD approach.

5. Experiments

5.1. Setting

Datasets. The generic distillation is conducted on ImageNet-1k [39] training set with 1.2M images. Following self-supervised recipes [13], we do not use the label information, so that lightweight models focus on soaking up the task-agnostic representations. In specific distillation, the models are fine-tuned from the previous stage on ImageNet-1k [39], MS COCO [26] and ADE20K [58] datasets.

Implementation details. In generic distillation stage, the MAE pre-trained ViT-Base model [13] is employed as the teacher. The student model is trained for 300 epochs using the AdamW optimizer [30], learning rate 2.4e-3, weight decay 0.05, batch size 4096, and image resolution 224×224. Unless specified, the mask ratio is set to 75% and the student decoder contains 4 Transformer blocks with 128 and 256 dimensions for ViT-Tiny and ViT-Small, respectively.

In task-specific distillation stage, the student decoder is discarded while the encoder is utilized as backbone to extract feature for various tasks, as do in MAE [13]. We use the official or re-implemented MAE fine-tuned model as the teacher. To avoid deteriorating the general representations obtained from the previous stage, a layer decay schedule is adopted to train the student model for all downstream tasks.

For image classification, we take a fine-tuned ViT-base model as the teacher, which is officially released by MAE [13] and achieves 83.6% top-1 accuracy. Following DeiT [41] distillation recipe, we append a distillation token to the student model for token-based distillation and use the hard decision of the teacher as the distillation label. The student model is trained for 200 epochs.

For object detection and instance segmentation tasks, we follow the ViTDet [25] framework, where the official ViTDet-Base [25] model are used as the teacher. The Feature-Richness Score method [12] is adopted to stress important features that are distilled from the teacher to the student model. Student models are trained with batch size 64 for 100 epochs. The input image resolution is 1024×1024.

For semantic segmentation, we use UperNet [51] task layers and distill the model for 160K iterations. Due to the absence of officially released model weights, we fine-tune

Table 4. Ablation study on single-stage and two-stage distillation methods, where G2SD w/o S.D denotes **only** performing generic distillation (*i.e.*, without specific distillation) and MAE🔥 means performing task-specific distillation during fine-tuning phase of MAE [13].

Method	Params (M)	Throughout (Images/s)	Generic Distillation	Specific Distillation	ImageNet-1k Top-1 Acc (%)	MS COCO AP^{bbox}	AP^{mask}	ADE20k mIoU
Teacher: ViT-Base	86.57	1.0×	N/A	N/A	83.6	51.6	45.9	48.3
Student: ViT-Tiny								
MAE [13]	5.72	5.84×	✗	✗	75.2	37.9	34.9	36.9
MAE🔥 [13]	5.91	5.74×	✗	✓	75.9	43.5	39.0	42.0
G2SD w/o S.D (*ours*)	5.72	5.84×	✓	✗	76.3	44.0	39.6	41.4
G2SD (*ours*)	5.91	5.74×	✓	✓	**77.0**	**46.3**	**41.3**	**44.5**
Student: ViT-Small								
MAE [13]	22.05	2.62×	✗	✗	81.5	45.3	40.8	41.1
MAE🔥 [13]	22.44	2.58×	✗	✓	81.9	48.9	43.5	44.9
G2SD w/o S.D (*ours*)	22.05	2.62×	✓	✗	82.0	49.9	44.5	46.2
G2SD (*ours*)	22.44	2.58×	✓	✓	**82.5**	**50.6**	**44.8**	**48.0**

the MAE pre-trained ViT-Base model on ADE20k by using the BEiT [4] semantic segmentation codebase to get teacher model, which achieves 48.3 mIoU, is comparable to MAE official report. During specific distillation, besides the supervision from the ground-truth, activation maps from the student and the teacher are aligned $w.r.t.$ the channel dimension [40].

5.2. Main Results

Image Classification. In Tab. 1, G2SD is compared with 1) supervised methods including MobileNet-v3 [19], ResNet [15, 48], DeiT [41, 42], Swin Trasnformer [28] and ConvNeXt [29]; 2) self-supervised methods upon ViT-Small, like BEiT [4] and CAE [8]; and 3) distillation methods upon vanilla ViTs, like DeiT🔥 [41], DearKD [7], Manifold [21], MKD [27], SSTA [49] and DMAE [3]. G2SD achieves 82.5% top-1 accuracy, which outperforms CNN-based ConvNeXt by 0.4%, by using fewer parameters (22M *vs.* 29M). G2SD consistently outperforms self-supervised methods, BEiT and CAE, by 0.8% and 0.5%, respectively. Compared with those distillation methods, G2SD shows the superiority. Remarkably, with the limited parameters (~6M), G2SD reports a substantial gain compared to DeiT-Ti🔥 and carefully designed MobileNet-v3.

Object Detection and Instance Segmentation. In Tab. 2, we report AP^{bbox} for object detection and AP^{mask} for instance segmentation. We compare G2SD with some popular methods on various backbone network: 1) vanilla ViT, like CAE [8], ViT-Adapter [9], imTED [56]; 2) elaborately designed architecture, like CNN based ConvNeXt [29] and hierarchical designed Swin Transformer [28]. One can see that G2SD-S, with fewer parameters, obtains more than 4.4 AP^{bbox} gains compared with ConvNeXt-T and Swin-T, which contain many inductive bias. Compared with CAE-S, which benefits from masked image modeling, G2SD also show the extraordinary superiority. Moreover, G2SD-S sig-

Table 5. Ablation study on generic distillation targets. e_i^t, $\hat{z}_i^t$ and $\mathcal{R}(e_i^t)$ respectively denote teacher encoder features, teacher decoder features, the relation among teacher encoder features. #5 is the default setting.

Target	e_i^t $i \in \mathcal{V}$	$\mathcal{R}(e_i^t)$ $i \in \mathcal{V}$	$\hat{z}_i^t$ $i \in \mathcal{V}$	$\hat{z}_i^t$ $i \in \mathcal{M}$	Accuracy (%)	mIoU (%)
#1	✓				81.60	43.69
#2		✓			81.45	43.64
#3				✓	81.96	45.20
#4	✓			✓	81.85	44.12
#5			✓	✓	**81.99**	**46.19**

nificantly outperforms imTED-S by 2.6 AP^{bbox} on object detection and 2 AP^{mask} on instance segmentation, where imTED-S uses pre-trained MAE encoder as backbone and pre-trained MAE decoder as task layers.

Semantic Segmentation. In Tab. 3, G2SD is compared with ViT-Adapter [9], ConvNeXt [29] and Swin Transformer [28]. G2SD-S outperforms all the compared methods by at least 1.4 mIoU, where ViT-Adapter elaborately modifies the model architecture for adapting dense prediction tasks. Remarkably, only using 11M parameters, G2SD-Ti achieves 44.5 mIoU, which pushes the performance of lightweight ViT models to a new height.

5.3. Ablation Studies: Single-stage *vs*. Two-stage

In Tab. 4, comprehensive experiments are conducted to compare single-stage and two-stage distillation methods. The teacher models include: 1) pre-trained MAE ViT-Base model for generic distillation; 2) fine-tuned MAE models on ImageNet-1k, MS COCO and ADE20k for specific distillation, which respectively reach 83.6% top-1 accuracy, 51.6 AP^{bbox}, 45.9 AP^{mask} and 48.3 mIoU. The student models are vanilla ViT-Tiny and ViT-Small, which are initialized from self-supervised method MAE [13] and G2SD. We denote MAE🔥 as model pre-trained with MAE and fine-tuned

with task specific distillation. MAE ViT-Small model is pre-trained for 300 epochs, by using the official codebase.

When specific distillation is not used, G2SD w/o S.D outperforms MAE by a large margin, *e.g.*, 49.9 *vs.* 45.3 AP^{bbox} and 46.2 *vs.* 41.1 mIoU, which benefits from the transferred task-agnostic knowledge. After activating specific distillation, both MAE⚗ and G2SD boost their performances, *e.g.*, G2SD-S achieves 0.7 AP^{bbox} gains and 1.8 mIoU gains, which are attributed to discriminative representation configuration. In conclusion, G2SD outperforms MAE and MAE⚗ across model sizes and datasets, validating the superiority of our two-stage distillation approach.

5.4. Ablation Studies: Generic Distillation

Target Configuration. We investigate the impact of target feature selection in generic distillation stage and report the results in Tab. 5. All models are trained under the same recipe and evaluated on ImageNet-1k and ADE20k. From Tab. 5 (#5), one can see that aligning student decoder features with teacher decoder's hidden features at both visible and masked patches achieves the best results, *e.g.*, 81.99% top-1 accuracy on ImageNet-1k and 46.19 mIoU on ADE20k.

Transferring from teacher encoder to student encoder is the most straightforward method, as shown in Tab. 5 (#1), but it only reaches 43.69 mIoU on ADE20k. The reason lies on that it overlooks the context understanding ability, which is beneficial for dense prediction tasks. Distilling the relation among tokens is popular and effective in NLP [46]. We thus conduct experiments using the self-attention relation of teacher encoder as distillation target, and find that the student only obtains 81.45% top-1 accuracy on ImageNet-1k, Tab. 5 (#2).

In Tab. 5 (#3), we align the student decoder features with those of the teacher decoder on the masked positions. In this way, the student respectively gets 0.36% accuracy and 1.51 mIoU gains on ImageNet and ADE20k compared to Tab. 5 (#1), which verifies the superiority of learning the context understanding capacity. Furthermore, we simultaneously calculate alignment loss on encoder features at visible patches and on decoder features at masked patches in Tab. 5 (#4), which is a more direct approach to let student inherit the feature extracting and context understanding capability of teacher, compared with Tab. 5 (#5). Unfortunately, the student performs worse than only calculating alignment loss on decoder features at masked patches, *e.g.*, 44.12 *vs.* 46.19 mIoU on ADE20k.

Mask Ratio. A high mask ratio (75%) works well in MAE [13], but the suitable mask ratio in generic distillation still needs to be explored. In general, predicting masked features is more challenging than predicting pixels. However, the observations are consistent with the teacher MAE,

Table 6. Ablation on the mask ratio (*top*) and target layer of the teacher model used for distillation (*bottom*).

Mask ratio	0.05	0.25	0.55	0.75	0.9
Top-1 Acc(%)	81.7	81.7	81.6	**82.0**	81.8
Layer Index	1	2	4	6	8
Top-1 Acc(%)	81.6	81.8	**82.0**	81.8	81.7

Table 7. Ablation study on the width and depth (D) of the student decoder. The depth and width of the teacher's decoder are 8 and 512, respectively.

Width	D	Acc(%)	D	Acc(%)	D	Acc(%)
128		81.9		81.8		81.7
256	2	81.7	4	**82.0**	8	81.7
512		81.8		81.7		80.3

as illustrated in Tab. 6 (*top*), where a high mask ratio tends to generate good results. The reason may be that the teacher model can express itself to the greatest extent when the mask ratio is consistent with the MAE pre-training phase.

Target Layer. A sufficiently deep decoder is essential for the fine-tuning performance in MAE [13]. We study the impact of which decoder layer is the best target layer Tab. 6 (*bottom*). One can see that using features of the 4-th layer as distillation targets for G2SD yields better accuracy. This can be explained that the last several layers in decoder are more specialized for low-level information (*e.g.*, pixel values) reconstruction while the first several layers in a decoder can't produce enough general representations.

Decoder Design. We study how the performance varies with decoder depth and width, where depth and width respectively denote the number of Transformer blocks and the embedding dimension of each Transformer block. As demonstrated in Tab. 7, the student decoder of width 256 and depth 4 yields optimal results, in terms of image classification. When the student decoder is heavy, the student encoder can be "lazy" to pursue good features as the decoder is competent for both feature extraction and image reconstruction.

5.5. Analysis

G2SD is compared with MAE⚗ and DeiT⚗ in terms of occlusion invariance, representation similarity and robustness, which indicate that it learns representations general to downstream tasks. DeiT⚗ denotes performing task-specific distillation by replacing the original teacher with the fine-tuned MAE-Base model. For fair comparison, we set the total training epochs of the three methods to be same (500
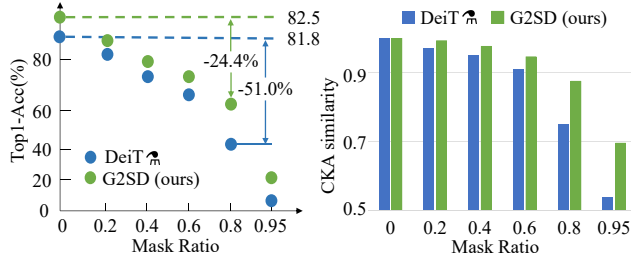
Figure 3. Performance degradation (*left*) and CKA similarity (*right*) between the representations generated by the complete image and the corrupted image with various mask ratios.

Table 8. Robustness evaluation. "IN" is short for ImageNet.

Methods	IN	IN-A	IN-R	IN-S	IN-V2
Teacher: ViT-Base	83.6	35.9	48.3	34.5	73.2
Student: ViT-Tiny					
DeiT🔥 [41]	75.3	9.5	36.2	23.4	63.3
MAE🔥 [13]	75.9	10.9	38.7	**26.3**	64.7
G2SD (ours)	**77.0**	**12.9**	**39.0**	25.9	**65.6**
Student: ViT-Small					
DeiT🔥 [41]	81.8	24.2	45.9	32.1	71.1
MAE🔥 [13]	81.9	26.6	**46.8**	**34.3**	71.1
G2SD (ours)	**82.5**	**29.4**	**46.8**	33.6	**72.1**

epochs). The major difference between those three methods is initialization, *i.e.*, G2SD is initialized from generic distillation, MAE🔥 from MAE pre-training, and DeiT🔥 from scratch.

Centered Kernel Alignment [24] (CKA) is a preferred metric evaluating normalized similarity between two feature maps or representations, and it is invariant to the orthogonal transformation of representations and isotropic scaling. We calculate CKA scores to analyze the occlusion invariance and representation similarity in the following.

Occlusion Invariance. Masked autoencoders are verified to learn occlusion invariant features in [23]. In Fig. 3 (*left*), we directly evaluate the performance of DeiT🔥 and G2SD under various mask ratios on the ImageNet-1k validation set. G2SD decreases about 24% while DeiT🔥 decreases about 51% when the mask ratio is 80%. In Fig. 3 (*right*), we calculate the CKA similarity between masked image representations and complete image representations, and find that G2SD can obtain higher CKA scores than DeiT🔥. These observations suggest that G2SD preserves more occlusion invariance than the single-stage method (*e.g.*, DeiT🔥).

Representation Similarity. In Fig. 4 (a) and (b), representations generated by G2SD w/o S.D is more similar with pre-trained MAE-B than pre-trained MAE-S, indicat-
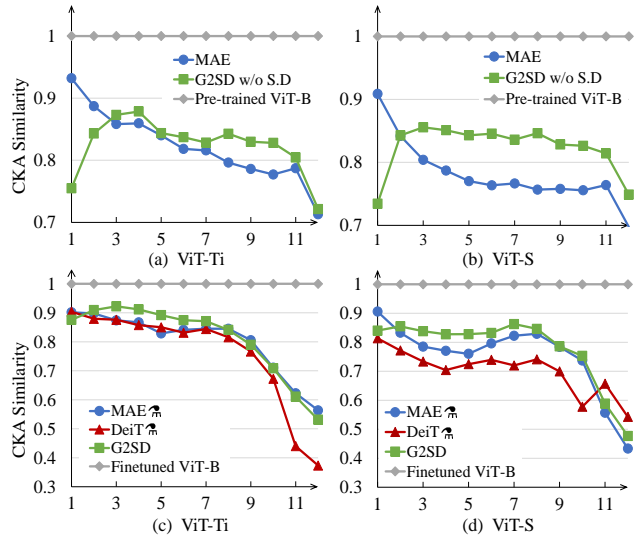

Figure 4. CKA similarity between representations generated by pre-trained MAE-B with: (a) pre-trained MAE-Ti and G2SD-Ti w/o S.D, and (b) MAE-S and G2SD-S w/o S.D. CKA similarity between representations generated by fine-tuned MAE-B with: (c) fine-tuned MAE-Ti🔥, DeiT-Ti🔥, and G2SD-Ti, and (d) fine-tuned MAE-S🔥, DeiT-S🔥, and G2SD-S. *x-axis* denotes network depth.

ing that generic distillation enables better features than simply reconstructing pixels. Furthermore, after task-specific distillation, G2SD consistently obtains higher CKA scores than MAE🔥 and DeiT🔥, as illustrated in Fig. 4 (c) and (d), implying that generic distillation provides a favored initialization for specific distillation.

Robustness. This is evaluated by testing the trained classifier on several ImageNet variants including ImageNet-A [17], ImageNet-R [16], ImageNet-S [45] and ImageNet-V2 [37]. From Tab. 8, one can see that G2SD outperforms the compared methods, which implies better generalization capability. In other words, the proposed G2SD encourages the small student model to maintain the generalization capability of teacher model endowed by the generic self-supervised method, as much as possible.

6. Conclusion

We proposed a two-stage distillation approach, termed generic-to-specific distillation (G2SD), to tap the potential of lightweight ViTs under the supervision of pre-trained large models. For generic distillation, we further designed a simple-yet-effective distillation strategy by aligning students' predictions with latent features of large masked autoencoders at both masked and visible patches. With two-stage distillation, the task-agnostic and task-specific knowledge of large models were transferred to lightweight ones. Extensive experiments on image classification, object de-

tection, and semantic segmentation validated the performance of the proposed G2SD approach, with striking contrast with state-of-the-art methods. This study has built a solid baseline for the two-stage vision model distillation.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas C. Damianou, Neil D. Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. *IEEE CVPR*, pages 9155–9163, 2019. 4

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 1, 2

[3] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Loddon Yuille, Yuyin Zhou, and Cihang Xie. Masked autoencoders enable efficient knowledge distillers. *ArXiv*, abs/2208.12256, 2022. 5, 6

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 1, 2, 5, 6

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 2, 5, 11

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Distilling knowledge via knowledge review. *IEEE CVPR*, pages 5006–5015, 2021. 2

[7] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. *IEEE CVPR*, pages 12042–12052, 2022. 2, 5, 6

[8] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 1, 5, 6

[9] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Y. Qiao. Vision transformer adapter for dense predictions. *ArXiv*, abs/2205.08534, 2022. 5, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019. 2

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *preprint arXiv:2010.11929*, 2020. 1, 2, 3

[12] Zhixing Du, Rui Zhang, Ming-Fang Chang, Xishan Zhang, Shaoli Liu, Tianshi Chen, and Yunji Chen. Distilling object detectors with feature richness. In *NeurIPS*, 2021. 3, 5, 11

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE CVPR*, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 11

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE ICCV*, 2017. 2, 5, 11

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE CVPR*, 2016. 1, 5, 6

[16] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE ICCV*, 2021. 8

[17] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE CVPR*, 2021. 8

[18] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 1, 2, 11

[19] Andrew G. Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *IEEE ICCV*, pages 1314–1324, 2019. 1, 2, 5, 6

[20] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. Learning what and where to transfer. In *ICML*. PMLR, 2019. 2

[21] Ding Jia, Kai Han, Yunhe Wang, Yehui Tang, Jianyuan Guo, Chao Zhang, and Dacheng Tao. Efficient vision transformers via fine-grained manifold distillation. *ArXiv*, abs/2107.01378, 2021. 2, 5, 6

[22] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. *ArXiv*, abs/1909.10351, 2020. 1, 3

[23] Xiangwen Kong and Xiangyu Zhang. Understanding masked image modeling via learning occlusion invariant feature. *ArXiv*, abs/2208.04164, 2022. 8

[24] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. *ArXiv*, abs/1905.00414, 2019. 8

[25] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ArXiv*, abs/2203.16527, 2022. 2, 5, 11

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5

[27] Jihao Liu, Boxiao Liu, Hongsheng Li, and Yu Liu. Meta knowledge distillation. *ArXiv*, abs/2202.07940, 2022. 2, 5, 6

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 5, 6

[29] Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *IEEE CVPR*, pages 11966–11976, 2022. 1, 5, 6

[30] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[31] Sachin Mehta and Mohammad Rastegari. Mobilevit: Lightweight, general-purpose, and mobile-friendly vision transformer. *ArXiv*, abs/2110.02178, 2022. 2

[32] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *NeurIPS*, 34:23296–23308, 2021. 2

[33] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. *IEEE CVPR*, pages 3962–3971, 2019. 2

[34] Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dongsheng Li, and Zhaoning Zhang. Correlation congruence for knowledge distillation. *IEEE ICCV*, pages 5006–5015, 2019. 2

[35] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. BEiT v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1, 2

[36] Zhiliang Peng, Zonghao Guo, Wei Huang, Yaowei Wang, Lingxi Xie, Jianbin Jiao, Qi Tian, and Qixiang Ye. Conformer: Local features coupling global representations for recognition and detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[37] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*. PMLR, 2019. 8

[38] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *CoRR*, abs/1412.6550, 2015. 1, 2

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 2, 5

[40] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen. Channel-wise knowledge distillation for dense prediction*. *IEEE ICCV*, pages 5291–5300, 2021. 3, 6

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *preprint arXiv:2012.12877*, 2020. 1, 2, 5, 6, 8, 11

[42] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 6

[43] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. *IEEE ICCV*, pages 1365–1374, 2019. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeuralIPS*, 2017. 3

[45] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, pages 10506–10518, 2019. 8

[46] Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *FINDINGS*, 2021. 7

[47] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. *arXiv preprint arXiv:2112.09133*, 2021. 1, 2

[48] Ross Wightman, Hugo Touvron, and Herv'e J'egou. Resnet strikes back: An improved training procedure in timm. *ArXiv*, abs/2110.00476, 2021. 6

[49] Haiyan Wu, Yuting Gao, Yinqi Zhang, Shaohui Lin, Yuan Xie, Xing Sun, and Ke Li. Self-supervised models are good teaching assistants for vision transformers. In *ICML*. PMLR, 2022. 2, 5, 6

[50] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *ECCV*, 2022. 2

[51] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 2, 5

[52] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *IEEE CVPR*, pages 9653–9663, 2022. 1

[53] Zhendong Yang, Zhe Li, Yuan Gong, Tianke Zhang, Shanshan Lao, Chun Yuan, and Yu Li. Rethinking knowledge distillation via cross-entropy. *ArXiv*, abs/2208.10139, 2022. 2

[54] Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. Vitkd: Practical guidelines for vit feature knowledge distillation. *ArXiv*, abs/2209.02432, 2022. 2

[55] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *IEEE CVPR*, pages 1204–1213, 2022. 1, 2

[56] Xiaosong Zhang, Feng Liu, Zhiliang Peng, Zonghao Guo, Fang Wan, Xian-Wei Ji, and Qixiang Ye. Integral migrating pre-trained transformer encoder-decoders for visual object detection. *ArXiv*, abs/2205.09613, 2022. 2, 5, 6

[57] Xiaosong Zhang, Yunjie Tian, Wei Huang, Qixiang Ye, Qi Dai, Lingxi Xie, and Qi Tian. Hivit: Hierarchical vision transformer meets masked image modeling. *ArXiv*, abs/2205.14949, 2022. 1

[58] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *IJCV*, 127(3):302–321, 2019. 2, 5

[59] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training

with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 5, 11

A. Hyperparameters

A.1. Image Classification

For distillation, as in [41], we added a learnable distillation token, which is combined with the cls token to produce final predictions in the inference phase. In experiments, the data augmentation and optimizer follow the fine-tuning recipe of MAE [13], while the learning rate, training epochs and layer-wise learning-rate decay are specified. For models training from scratch (e.g., DeiT🐟), we set the layer decay value as 1.0, which means no layer decay is adopted. For pre-trained models (e.g., MAE [13], G2SD), we set the layer decay value to 0.75 and training epochs to 200.

Table 9. Hyperparameters for distilling on ImageNet-1K.

Hyperparameters	Value (Fine-tuning)	Value (From scratch)
Training epochs	200	500
Base learning rate	1e-3	2.5e-4
Layer decay	0.75	1.0
Warm up epochs	5	
Label smoothing	0.1	
Mixup	0.8	
Cutmix	1.0	
Drop path	0.0	
Batch size	1024	
Weight decay	0.05	
Optimizer	AdamW	
Learning rate schedule	Cosine decay	
Augmentation	RandAug(0,0.5)	
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$	

A.2. Object Detection and Instance Segmentation

In the experiments, we adopt the official codebase[2] and follow the settings used in ViTDe [25]. The total batch size is set to 64 (8 images per GPU). The learning rate is set to $1e^{-4}$, the backbone's drop path rate is 0.1, and the distill warm step is 500. The overall training target is the same as [12]: $L = L_{GT} + \alpha L_{FPN} + \beta L_{head}$, where α and β are respectivvely set to 0.001 and 0.1.

A.3. Semantic Segmentation

In this experiment, we adopt the BEiT's segmentation codebase[3] and set the total batch size to 32 (4 images per

---

[2] https : / / github . com / facebookresearch / detectron2/tree/main/projects/ViTDet

[3] https://github.com/microsoft/unilm/beit

GPU). The backbone's drop path rate is 0.1. The layer decay rate is 0.75. The learning rate of ViT-Small and ViT-Tiny are respectively set to $2e^{-4}$ and $5e^{-4}$. We set the temperature parameter $\tau = 1$, the loss weight $\alpha = 3$ for the logits map distillation.

B. Training Time and Efficiency

As shown in Table 10, G2SD outperforms DeiT [41] and DeiT🐟 [41], which have a longer training schedule (500 epochs). The teacher of DeiT🐟 is the same as G2SD's. In the generic distillation stage, since the input of G2SD is a masked image (75% patches are discarded), the training time per epoch is less than DeiT (which computes the whole image).

Table 10. G2SD vs DeiT. The total training epochs is 500.

Methods	1-st stage	2-nd stage	Time	Top-1 Acc (%)
G2SD	G.D 300 epochs	S.D 200 epochs	71 h	82.5
DeiT🐟	Supervised+Distillation 500 epochs		112 h	81.7 (-0.8)
DeiT	Supervised 500 epochs		53 h	81.4 (-1.1)

C. Detection Performance with ViTDet

For the lack of official Mask-RCNN [14] results and checkpoints of MAE [13], we choose ViTDet [25] as the detector. In Table 11, the backbone models are initialized from various supervisions, *e.g.*, supervised methods (DeiT [41]), distilled methods (DeiT🐟 [41] and G2SD) and self-supervised methods (DINO [5] and iBoT [59]). From Table 11, G2SD significantly outperforms competitors on performance and convergence speed.

Table 11. Performance on MS COCO using the ViTDet framework [25], which is trained for 100 epochs with single-scale input (1024×1024).

Methods (Supervision)	ImageNet Acc (%)	AP^{bbox}	AP^{mask}
DeiT-S (sup., 300e)	79.9	45.7	40.7
DeiT-S🐟 (sup.&distill., 300e)	81.2	47.2	41.9
DeiT-S (sup., 500e)	81.4	46.9	41.6
DINO-S (self-sup., 3200e)	82.0	49.1	43.3
iBOT-S (self-sup., 3200e)	82.3	49.7	44.0
G2SD-S (w/o S.D, 300e)	82.0	49.9	44.5
G2SD-S (300e)	82.5	50.6	44.8

D. More Ablations on Target Configuration

In Table 5, we have conducted ablation studies on intermediate features as generic distillation targets. Compared with using intermediate features as distillation targets, taking the teacher's prediction as distillation objective [18,41] is also a popular alternative. Therefore, we take

11

Table 12. Ablation study of distillation targets on ImageNet-1k. 'S.D' is short for specific distillation.

Distillation targets	W/O S.D Acc (%)	W S.D Acc (%)
Our default settings	**82.0**	**82.5**
MAE's reconstructions	81.4	81.8
MAE's reconstructions + GT	81.5	81.7

the MAE's predictions as the generic distillation targets in Table 12. When taking the MAE's predictions as the targets for masked positions, the performance drops to 81.4% (without specific distillation) and 81.8% (with specific distillation). This observation is consist with the results in Table 6 (*bottom*), where the last several layers in decoder are more specialized for low-level information reconstruction task.