CLIPの微調整性能の向上

Yixuan Wei[1], Han Hu[2]*, Zhenda Xie[1], Ze Liu[3], Zheng Zhang[2], Yue Cao[2],
Jianmin Bao[2], Dong Chen[2], Baining Guo[2]
[1]Tsinghua University [2]Microsoft Research Asia [3]USTC
{tt-yixuanwei,hanhu,t-zhxie,t-liuze,zhez,yuecao,jianbao,doch,bainguo}@microsoft.com。

Abstract

CLIPモデルは非常に高いゼロショット認識精度を示しているが、下流の視覚タスクに対する微調整性能は最適ではない。これとは対照的に、マスク画像モデリング（MIM）は、訓練中に意味ラベルがないにもかかわらず、下流タスクの微調整において優れた性能を発揮する。この2つのタスクは、画像レベルのターゲットとトークンレベルのターゲット、クロスエントロピー損失と回帰損失、全画像入力と部分画像入力という異なる要素を持つ。この違いを緩和するために、古典的な特徴マップ蒸留フレームワークを導入する。このフレームワークは、CLIPモデルの意味的能力を継承すると同時に、MIMの主要な要素を組み込んだタスクを構成することができる。実験により、特徴マップ蒸留アプローチは、いくつかの典型的な下流視覚タスクにおいて、CLIPモデルの微調整性能を大幅に向上させることが示唆された。また、このアプローチにより、MIMといくつかの診断特性を共有する新しいCLIP表現が得られることが確認された。さらに、特徴マップ蒸留アプローチは、DINO、DeiT、SwinV2-Gのような他の事前学習モデルにも一般化し、COCO物体検出において64.2mAPの新記録を達成し、+1.1の改善を示した。コードとモデルは https://github.com/ SwinTransformer/Feature-Distillation で公開されている。

1. Introduction

事前学習と微調整のパラダイムは、[20, 30, 13, 39]のような多くの影響力のある作品によって証明されているように、コンピュータビジョンにおける深層学習手法の成功に役立っている。一般的なプラクティスの1つは、物体検出[13]やセマンティックセグメンテーション[39]など、様々な下流のビジョンタスクの初期化として、ImageNet-1k分類タスク[10]で事前に訓練されたモデル重みを使用することである。しかし、このアプローチは2つの重要な課題に直面している：

* 共著者 この研究は、Yixuan Wei、Zhenda Xie、Ze Liu が Microsoft Research Asia でインターンをしているときに行われました。

表1：古典的な特徴蒸留フレームワークによるViT-B／16 CLIPモデル[42]の微調整性能の向上。このモデルはImageNet-1Kデータセット[10]で300エポックの画像のみで蒸留される。その結果、4つの評価ベンチマークで明確な改善が見られた。MAE[17]の結果も参考のためグレーで示す。

Method	IN-1K (%)		ADE20K	COCO		NYUv2
	linear	f.t.	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
MAE [17]	68.0	83.6	48.1	46.5	40.9	0.383
CLIP [42]	79.5	82.9	49.5	45.0	39.8	0.416
FD-CLIP	80.1	85.0	51.7	48.2	42.5	0.352
Δ	**↑0.6**	**↑2.1**	**↑2.2**	**↑3.2**	**↑2.7**	**↓0.064**

高品質の画像分類データをスケールアップすることの難しさ、およびカテゴリラベルに含まれる限定的な意味情報。
最近のCLIP [42]は、これらの課題を軽減している。これは、ウェブスケールのノイジーな視覚-言語ペアから表現を学習するために、対照学習を利用する。学習された表現 は、ゼロショット画像分類や画像-テキスト検索タスクの性能から明らかなように、 印象的な意味モデル化能力を示す。一方、マスク画像モデリング(MIM)[2, 58, 17]に基づく新しい自己教師付き事前学習法も、様々な下流タスクに対する優れた微調整性能により、大きな注目を集めている。本稿では、一般性を損なわない範囲で、主にMAE [17]について述べる。

2つの事前学習方法を比較すると、CLIPモデルはImageNet-1Kでの優れた線形プロービング性能に反映されるように、より豊富な意味情報を学習する。しかし、他のほとんどのタスクにおけるCLIPの微調整性能は、Tab. 1. 通常、より優れた意味情報を持つモデルは、より優れた移植性を持つと考えられているため、この観察は直感に反しているように見える。CLIPは微調整においてMIMと同じように、あるいはMIMを上回ることができるのだろうか？

Improving CLIP Fine-tuning Performance

Yixuan Wei[1], Han Hu[2]*, Zhenda Xie[1], Ze Liu[3], Zheng Zhang[2], Yue Cao[2],
Jianmin Bao[2], Dong Chen[2], Baining Guo[2]
[1]Tsinghua University [2]Microsoft Research Asia [3]USTC
{t-yixuanwei,hanhu,t-zhxie,t-liuze,zhez,yuecao,jianbao,doch,bainguo}@microsoft.com

Abstract

CLIP models have demonstrated impressively high zero-shot recognition accuracy, however, their fine-tuning performance on downstream vision tasks is sub-optimal. Contrarily, masked image modeling (MIM) performs exceptionally for fine-tuning on downstream tasks, despite the absence of semantic labels during training. We note that the two tasks have different ingredients: image-level targets versus token-level targets, a cross-entropy loss versus a regression loss, and full-image inputs versus partial-image inputs. To mitigate the differences, we introduce a classical feature map distillation framework, which can simultaneously inherit the semantic capability of CLIP models while constructing a task incorporated key ingredients of MIM. Experiments suggest that the feature map distillation approach significantly boosts the fine-tuning performance of CLIP models on several typical downstream vision tasks. We also observe that the approach yields new CLIP representations which share some diagnostic properties with those of MIM. Furthermore, the feature map distillation approach generalizes to other pre-training models, such as DINO, DeiT and SwinV2-G, reaching a new record of 64.2 mAP on COCO object detection with +1.1 improvement. The code and models are publicly available at https://github.com/SwinTransformer/Feature-Distillation.

1. Introduction

The pre-training and fine-tuning paradigm is instrumental in the success of deep learning methods in computer vision, as evidenced by numerous influential works such as [20, 30, 13, 39]. One common practice is to use model weights pre-trained on ImageNet-1k classification task [10] as the initialization for various downstream vision tasks, such as object detection [13] and semantic segmentation [39]. However, this approach faces two key challenges: the difficulty in scaling up high-quality image classification data, and the limited semantic information contained in category labels, both of which constrain the ability to further improve model performance.

Table 1: Improving the fine-tuning performance of the ViT-B/16 CLIP model [42] via a classical feature distillation framework. The model is distilled on ImageNet-1K dataset [10] with images only for 300 epochs. Clear gains are observed on four evaluation benchmarks. MAE [17] results are also listed in gray for reference.

Method	IN-1K (%)		ADE20K	COCO		NYUv2
	linear	f.t.	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
MAE [17]	68.0	83.6	48.1	46.5	40.9	0.383
CLIP [42]	79.5	82.9	49.5	45.0	39.8	0.416
FD-CLIP	80.1	85.0	51.7	48.2	42.5	0.352
Δ	↑**0.6**	↑**2.1**	↑**2.2**	↑**3.2**	↑**2.7**	↓**0.064**

The recent CLIP [42] alleviates these challenges. It utilizes contrastive learning to learn representations from web-scale noisy vision-language pairs. The learned representations exhibit impressive semantic modeling capabilities, as evidenced by performance on zero-shot image classification and image-text retrieval tasks. Meanwhile, a new self-supervised pre-training method based on masked image modeling (MIM) [2, 58, 17] has also attracted great attention for its excellent fine-tuning performance on various downstream tasks. Without losing generalizability, we mainly discuss MAE [17] in this paper.

When comparing the two pre-training methods, the CLIP model learns richer semantic information reflected by its superior linear probing performance on ImageNet-1K. However, its fine-tuning performance on most other tasks are worse than MAE, as shown in Tab. 1. This observation appears counter-intuitive since models with better semantics are usually considered to have better transferability. This raises a further question: can CLIP be made as successful as, or even surpass, MIM in fine-tuning? To answer this question, we firstly decompose the ingredients of these pre-training methods into three aspects: input ratios, training

*Corresponding Author. The work is done when Yixuan Wei, Zhenda Xie, and Ze Liu are interns at Microsoft Research Asia.

表2：ImageNet-1K分類におけるViTL／14 CLIPモデル[42]の微調整性能の向上。

Method	Res.	Pre-train datasets	IN-1K(%)
WiSE-FT [54]	336^2	WIT-400M [42]	87.1
DeiT III [52]	384^2	IN-22K [10]	87.7
ViT [11]	512^2	JFT-300M [50]	87.8
Scaling [63]	384^2	JFT-3B[63]	88.5
BeiT [2]	512^2	DALLE [45] & IN-22K	88.6
CLIP [42]	224^2	WIT-400M	86.1
FD-CLIP	224^2	WIT-400M	**87.7** (+1.6)
	224^2	WIT-400M & IN-22K	**88.3**
	336^2	WIT-400M & IN-22K	**89.0**

この問いに答えるため、まず、これらの事前学習法の構成要素を、表3に示すように、入力比率、学習対象の粒度、学習損失の3つの側面に分解する。3. CLIPと2つの典型的なMIMアプローチとの違いを比較することで、CLIPの微調整性能が劣っている原因として学習損失を除外し、入力比率（すなわち、全画像対部分画像）と学習対象粒度（すなわち、画像レベル対トークンレベル）が重要な要因ではないかと推測する。入力比率の差を縮めることは簡単であるが、CLIPの学習ターゲットの粒度を画像レベルからトークンレベルに変更することは、既存の視覚言語学習データが画像レベルの監視に適しており、きめ細かな情報が不足しているため、大きな課題となる。

知識蒸留[21]は、あるモデルから別のモデルへ情報を転送するために広く 用いられている手法であり、一般的にはモデル圧縮を目的としている。本論文では、意味情報を保持したまま、CLIPモデルの学習対象粒度を画像レベルからトークンレベルに変換するためのブリッジとしても蒸留が機能することを実証する。具体的には、事前に訓練されたCLIPモデルを教師モデルとし、その出力特徴マップを蒸留対象として、この情報を、教師モデルと同じアーキテクチャとサイズを共有するランダムに初期化された生徒モデルに蒸留する。このプロセスは図1に示されており、ロジット蒸留[21，51，12]と区別するために「特徴蒸留」と呼ぶ。注目すべきは、生徒が教師モデルの出力を模倣しているにもかかわらず、両者の最適化パスが異なるため、中間層で異なる診断特性が得られることである。

蒸留フレームワークの柔軟性により、適切な帰納的バイアスと正則化を導入することで、生徒モデルの最適化経路を形成し、下流のタスクにおける生徒モデルの性能を向上させることができる。具体的には、いくつかの重要な調整を提案する： 1）教師特徴マップの標準化。この調整により、教師モデルに含まれる微妙な情報が増幅され、出力値の範囲が安定する。

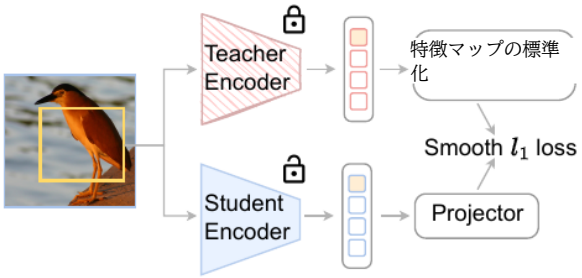


図1：トークンレベルのターゲットを導入し、事前学習されたCLIPモデルを蒸留する特徴マップ蒸留の説明図。オレンジ色のブロックは[CLS]トークンを表し、オレンジ色のボックスは元画像のランダムな切り出しを意味する。

この非対称な正則化により、生徒の表現のロバスト性が強化され、正確で一貫性のある教師信号が得られます。3）相対位置バイアスの共有。この帰納的バイアスの導入により、生徒モデルの並進不変性がさらに強化される。
特徴抽出フレームワークと上記の調整により、CLIPの強い意味情報を保持しつつ、下流のタスク微調整に優しいモデルを導出する。1. 様々なタスクにおいて、オリジナルのCLIPと比較して一貫した改善が見られる： ImageNet-1K画像分類[10]では+2.1の精度向上、ADE20Kセマンティックセグメンテーション[66]では+2.2mIoUの向上、COCOオブジェクト検出とインスタンスセグメンテーション[36]では+3.2ボックスAPと+2.7マスクAP、NYUv2深度推定[49]ではRMSE(↓)を0.064削減。この改善は、表2に示すように、ImageNet-1Kで+1.6の精度向上と、最大のCLIP-L／14モデルにスケールアップしても維持される。2. さらに、DINO [3]、DeiT [51]、先進的なSwinV2-G [37]のような他のモデルに我々の手法を一般化しても、様々なダウンストリームタスクで明確な利益を得ることができ、特にCOCO物体検出で64.2mAPの新記録を達成し、SwinV2-Gで+1.1mAPの改善が見られた。

これらの実験結果の改善に加えて、我々はさらに、異なるモデルから学習された視覚表現の特性を分析するためのいくつかの診断ツールを導入する。これらの分析により、特徴抽出がCLIPモデルをどのように改善するかを理解するための深い洞察が得られる： 1）CLIPモデルの異なる注目ヘッドをより深い層で多様化する、2）学習された表現の並進不変性を改善する、3）損失ランドスケープを平坦化し、最適化しやすさを反映する。

我々の貢献は以下の通りである：

- CLIPとMIM手法の成分の違いを検証し、MIMが微調整に成功するためには、ターゲットの粒度が重要であることを示す。
- 我々は古典的な特徴マップ蒸留を利用して、CLIPの学習ターゲットの粒度をトークンレベルのものに変換する。

Table 2: Improving the fine-tuning performance of the ViT-L/14 CLIP model [42] on ImageNet-1K classification.

Method	Res.	Pre-train datasets	IN-1K(%)
WiSE-FT [54]	336^2	WIT-400M [42]	87.1
DeiT III [52]	384^2	IN-22K [10]	87.7
ViT [11]	512^2	JFT-300M [50]	87.8
Scaling [63]	384^2	JFT-3B[63]	88.5
BeiT [2]	512^2	DALLE [45] & IN-22K	88.6
CLIP [42]	224^2	WIT-400M	86.1
FD-CLIP	224^2	WIT-400M	**87.7** (+1.6)
	224^2	WIT-400M & IN-22K	**88.3**
	336^2	WIT-400M & IN-22K	**89.0**

target granularity and training losses, as listed in Tab. 3. By comparing the differences between CLIP and two typical MIM approaches, we exclude the training losses to be responsible for the inferior fine-tuning performance of CLIP, and speculate that the input ratios (*i.e*. full image *vs*. partial image) and training target granularity (*i.e*. image-level *vs*. token-level) might be key factors. Although narrowing the differences in input ratios is straightforward, changing the granularity of the CLIP training targets from image-level to token-level poses a significant challenge, since existing vision-language training data is more suitable for image-level supervision and lacks fine-grained information.

Knowledge distillation [21] is a widely used technique for transferring information from one model to another, typically for the purpose of model compression. In this paper, we demonstrate that distillation can also perform as a bridge for converting the training target granularity of CLIP models from image-level to token-level, while preserving the semantic information. Specifically, we take the pre-trained CLIP model as the teacher model, use its output feature map as the distillation target, and distill this information into a randomly initialized student model that shares the same architecture and size as the teacher model. This process is illustrated in Fig. 1, which we refer to as *"feature distillation"* to differentiate it from logits distillation [21, 51, 12]. Notably, although the student mimics the teacher model's output, their different optimization paths can lead to different diagnostic properties on the inter-mediate layers, which is thought to be important for fine-tuning.

The flexibility of the distillation framework allows us to introduce proper inductive bias and regularization to shape the optimization path of the student model and enhance the student model performance on downstream tasks. Specifically, we propose several crucial adjustments: 1) Standardization of the teacher feature map. This adjustment amplifies the subtle information contained within the teacher model and stabilizes the output value range; 2) Asymmetric drop path rates for the teacher and student models. This asymmetric regularization enhances the robustness of

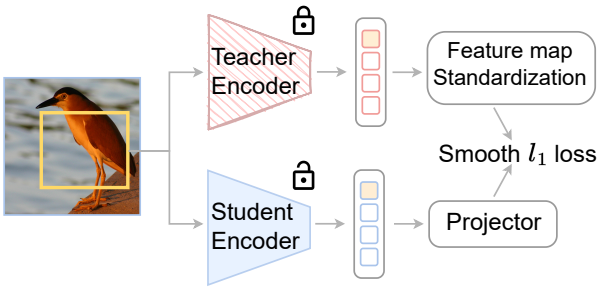


Figure 1: Illustration of the feature map distillation that introducing token-level targets to distill the pre-trained CLIP models. The orange block stands for `[CLS]` token, and the orange box means a random crop of the original image.

student representations and results in accurate and consistent teacher signals. 3) Shared relative position bias. The introduction of this inductive bias further strengthens the translation-invariant property of the student model.

With the feature distillation framework and the above adjustments, we derive a model that preserves the strong semantic information of CLIP while being friendly to downstream task fine-tuning, as shown in Tab. 1. We observe consistent improvements compared to the original CLIP on various tasks: +2.1 accuracy gains on ImageNet-1K image classification [10], +2.2 mIoU gains on ADE20K semantic segmentation [66], +3.2 box AP and +2.7 mask AP on COCO object detection and instance segmentation [36], and reducing RMSE(↓) by 0.064 on NYUv2 depth estimation [49]. The improvement remains when scaling up to the largest CLIP-L/14 model with a +1.6 accuracy gain on ImageNet-1K, as shown in Tab. 2. Moreover, when generalizing our method to other models, like DINO [3], DeiT [51] and the advanced SwinV2-G [37], we still earn clear gains on various downstream tasks, especially reaching a new record of 64.2 mAP on COCO object detection with +1.1 mAP improvement on SwinV2-G.

In addition to these improvements in experimental results, we further introduce several diagnostic tools to analyze the properties of learned visual representations from different models. These analyses provide deeper insights into understanding how feature distillation improves the CLIP model: 1) diversifying different attention heads of the CLIP model in deeper layers; 2) improving the translational invariance of learned representations; and 3) flattening the loss landscapes and reflecting optimization friendliness.

Our contributions are summarized as follows:

- We examine the ingredient differences between CLIP and MIM methods and demonstrate target granularity is vital in the success of MIM in fine-tuning.
- We leverage the classical feature map distillation to convert the training target granularity of CLIP to token-level ones, which enhances its fine-tuning per-

を、その意味情報を保持したまま形成する。

- 我々は、標準化された特徴マップの抽出、非対称なドロップパス率、共有された 相対位置バイアスなど、改善をさらに拡大する特徴抽出中のいくつかの重要な技 術を提案する。
- いくつかの診断ツールを用いて、CLIPと比較して、MIMとFD-CLIPの両方が、直感的に良いいくつかの特性を持っていることがわかった。
- 我々は、様々な事前学習モデルに我々の方法を一般化し、一貫した利得を観察した。また、先進的な3B SwinV2-Gモデルを我々のフレームワークで改良することで、COCO物体検出の新記録を樹立した。

2. Related Work

表現学習 視覚分野では、4つの注目すべき表現学習アプローチがある。1)教師付きデータセット[10, 50]上での画像分類(CLS)は、AlexNet[30]以来10年近く、上流での標準的な事前学習タスクである。事前学習された重みは、画像セグメンテーション[66, 16]、物体検出[36, 46]、ビデオ認識[24, 15]など、数多くのダウンストリームタスクに適用されている。2) 対照的言語画像事前学習(CLIP)タスクは、対になった画像とテキストを連結し、対になっていないものを分離するタスクであり、ゼロショット認識の分野を開拓し[42, 23]、マルチモダリティのダウンストリームタスク[47, 40, 44]で威力を発揮する。3)インスタンス対比学習(CLR)法は、同じ画像の補強ビューを他と対比することで、自己教師ありの方法で事前学習を行う[18, 57]。この方法は、線形評価と少数ショット評価を用いて、印象的な精度を達成する[3, 5, 31]。4) マスキング画像モデリング(MIM)もまた、自己教師ありの方法で表現を学習し、最初に画像領域の大部分をマスクし、マスクされた領域の画素値や特徴を予測するように学習する。これは微調整評価に優れている[2, 17, 58]。

本論文では、古典的な特徴マップ蒸留の枠組みを採用し、下流タスクでより良い性能を発揮し、元のCLIPに組み込まれた意味情報をほぼ保持する、同じサイズのリフレッシュCLIPモデルを導出することを提案する。また、DeiT[51]やSwinV2[37]のような教師付きモデルや、DINO[3]のような自己教師付きモデルを含む、非CLIPモデルにも我々の手法を一般化する。

知識蒸留 知識蒸留[21, 26]はCNNモデルで最初に提案され、教師付き学習性能を向上させ、コンパクトで小さなモデルを圧縮するためにTransformers[51, 53]でさらに研究されている。「暗い知識」は、生徒モデルが教師のロジット予測を模倣するとき、蒸留において有用であることが証明されている[12]。

表3: 入力比率、学習対象の粒度、損失形式の観点からCLIP法とMIM法の成分比較を行う。

Method	Input	Target	Loss	Semantics
BeiT [2]	Partial	Token-level	Cross-entropy	
MAE [17]	Partial	Token-level	Regression	
CLIP [42]	Full	Image-level	Cross-entropy	✓
FD-CLIP	Full	Token-level	Regression	✓

教師あり学習だけでなく、知識蒸留は深層強化学習[48, 4]、生涯学習[62]、推薦システム[7]でも広く使われている。我々の研究は、新しい蒸留アプローチを提案することを目的とするものではなく、CLIPとMIM手法の間の入力比率とターゲット粒度の違いを緩和するために、重要な設計でこの柔軟なフレームワークを活用する。我々は、トークン・レベルのターゲット粒度を持つタスクが、MIM法の成功に重要な要素であることを発見した。また、特徴抽出はCLIPのカウンターパートとして機能し、CLIPの微調整性能を向上させ、元の意味情報をほぼ維持する。

モデルの診断と説明 モデルの診断は、ディープラーニング・モデルの「ブラックボックス」を解明するために重要である。また、注意分析[67, 11, 56]、損失ランドスケープの可視化[33]、CKA[28]、知識ニューロンの発見[9, 14]など、トランスフォーマーを理解しようとする研究[67, 11, 9, 14, 29, 43, 41]もある。これらの研究に触発され、我々はMIM法のユニークな特性を明らかにするために、一連の注意と最適化関連の診断ツールを採用する。また、これらのツールをFD-CLIPにも適用することで、特徴抽出プロセスのより良い理解を提供する。

3. 特徴抽出によるCLIPの改善

CLIP法は、膨大な画像とテキストのペアを対比することで学習される豊富なセマンティクスを取り込む能力で知られている。MIM手法（例えばMAE）と比較して、CLIPは、その優れた線形プロービング性能（表1に示すように）によって証明されるように、より人間の概念に一致している。しかし、その印象的なセマンティック能力は、下流のタスクの微調整にはわずかな利益しかもたらさないようである。図2に示す物体検出タスクの学習損失曲線を精査すると、分類損失、すなわちL_{cls}は、MAEの事前学習とCLIPの事前学習で近い値を示すが、MAEの方がL_{bbox}が低く、より優れた定位能力を持つように見える。これらの違いは、CLIPの最適な微調整性能の背後にある要因を調査し、CLIPの強力なセマンティック能力をより発揮させる方法を探る動機となった。

表3に示すCLIP法とMIM法の成分の違いを比較する。3. それに基づいて、以下のように仕様化する。

formance and preserves its semantic information.

- We propose several crucial techniques during feature distillation that further enlarge the improvements, including distilling standardized feature maps, asymmetric drop path rates, and shared relative position bias.
- With several diagnostic tools, we find that compared to CLIP, both MIM and **FD**-CLIP possess several properties that are intuitively good, which may provide insights on their superior fine-tuning performance.
- We generalize our method to various pre-training models and observe consistent gains. We also set a new record on COCO object detection, by improving the advanced 3B SwinV2-G model with our framework.

2. Related Work

Representation Learning There are four notable representation learning approaches in vision area. 1) Image classification (CLS) on supervised datasets [10, 50] has been the standard upstream pre-training task for nearly a decade since AlexNet [30]. The pre-trained weights are applied to numerous down-stream tasks including image segmentation [66, 16], object detection [36, 46] and video recognition [24, 15]. 2) The contrastive language-image pre-training (CLIP) task is to connect paired images and texts and separate unpaired ones, which opens up the field of zero-shot recognition [42, 23], and proves to be powerful in multi-modality down-stream tasks [47, 40, 44]. 3) Instance contrastive learning (CLR) method performs pre-training in a self-supervised manner by contrasting the augmentation views of the same image with others [18, 57]. The method achieves impressive accuracy using linear and few-shot evaluations [3, 5, 31]. 4) Masked image modeling (MIM) learns representations also in a self-supervised way, which first masks a large portion of the image area and learns to predict the pixel values or features of the masked area. It excels in fine-tuning evaluations [2, 17, 58].

In this paper, we propose to adopt the classical feature map distillation framework to derive a same-size refreshed CLIP model which performs better on downstream tasks and largely preserves the semantic information incorporated in original CLIP. We also generalize our method to non-CLIP models, including supervised models, like DeiT [51] and SwinV2 [37] and self-supervised models, like DINO [3].

Knowledge Distillation Knowledge distillation [21, 26] is firstly proposed in CNN models, and further explored in Transformers [51, 53] to boost the supervised training performance and compress a compact small model. *"Dark knowledge"* is proven to be useful in distillation when student models mimic the logits prediction of the teachers [12]. Beyond supervised learning, knowledge distillation is also widely used in deep reinforcement learning [48, 4], lifelong learning [62] and recommendation system [7].

Our work does *not* aim to propose a new distillation approach, but leverage this flexible framework with crucial designs to mitigate the differences on input ratios and target granularity between CLIP and MIM methods. We find that tasks with token-level target granularity maybe a key ingredient to the success of MIM methods. And feature distillation serves as a counter-part for CLIP, which improves its fine-tuning performance and largely maintain its original semantic information.

Table 3: Ingredients comparison between CLIP and MIM methods from the perspective of input ratios, training target granularity and loss format.

Method	Input	Target	Loss	Semantics
BeiT [2]	Partial	Token-level	Cross-entropy	
MAE [17]	Partial	Token-level	Regression	
CLIP [42]	Full	Image-level	Cross-entropy	✓
FD-CLIP	Full	Token-level	Regression	✓

Model Diagnosing and Explanation Model diagnosing is important for demystifying the "black box" of deep learning models due to their high-dimensional and non-linear nature [59]. There have been also works [67, 11, 9, 14, 29, 43, 41] seeking to understand Transformers, including attention analysis [67, 11, 56], loss landscapes visualization [33], CKA [28] and knowledge neuron discovery [9, 14]. Inspired by these works, we adopt a set of attention- and optimization-related diagnostic tools to reveal the unique properties of MIM method. And these tools are also applied on **FD**-CLIP to provide a better understanding of feature distillation process.

3. Improving CLIP by Feature Distillation

The CLIP method is known for its ability to incorporate rich semantics learned by contrasting tremendous image-text pairs. Compared to the MIM methods (*e.g.* MAE), CLIP is more consistent with human concepts, as evidenced by its superior linear probing performance (as shown in Tab. 1). However, its impressive semantic capability seems to marginally benefit downstream tasks fine-tuning. By closely examining the training loss curves for the object detection task, depicted in Fig. 2, we note that the classification loss, *i.e.* L_{cls}, is close between MAE pre-training and CLIP pre-training, but MAE appears to have better localization ability with lower L_{bbox}. These differences motivated us to investigate the factors behind CLIP's sub-optimal fine-tuning performance and explore ways to unleash its powerful semantic capability better.

We compare the ingredient differences between CLIP and MIM methods shown in Tab. 3. Based on it, we spec-

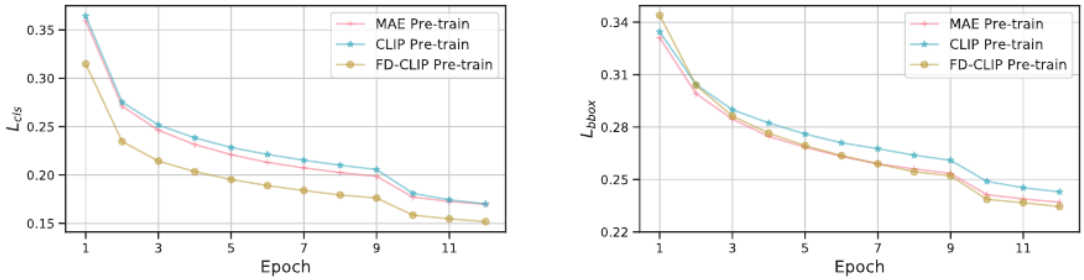


図2：COCO物体検出タスクにおけるMAE、CLIP、FD-CLIPの微調整。L_{cls}とL_{bbox}の学習エポックに対する損失曲線を可視化した。L_{cls}については、CLIPによる事前学習はMAEによる事前学習と同等であるが、L_{bbox}の曲線に反映されるように、局所化能力は劣る。

入力の比率（すなわち、フル画像対部分画像）と学習ターゲットの粒度（すなわち、画像レベル対トークンレベル）がCLIPの劣った微調整性能の原因かもしれない。CLIPで部分入力を使用することは比較的簡単ですが、CLIPや既存の視覚言語学習データは画像レベルの監視用に設計されているため、学習対象を画像レベルからトークンレベルに直接変更することは困難です。さらに重要なことは、CLIPの再トレーニングはコストがかかるため、避けたい。

そこで、CLIPモデルの学習対象粒度を画像レベルからトークンレベルに変換するためのブリッジとして、通常モデル圧縮に用いられる蒸留技術を、事前学習済みモデルの意味情報を保持したまま利用することを提案する。具体的には、図1に示すように、事前に訓練されたモデルが凍結された教師として機能し、ランダムに初期化された重みを持つ新しい同じモデルが生徒として機能する。

この手法では、従来の蒸留[21, 51]のようにロジットを蒸留するのではなく、「特徴蒸留」と呼ばれる、事前学習されたモデルの完全な出力特徴マップを蒸留対象として採用する。このアプローチにより、ロジット出力を持たないモデルも含め、どのような事前学習済みモデルでも扱うことができる。さらに、特徴マップを蒸留することで、単一の特徴ベクトルだけを蒸留するよりも高い微調整精度が得られ（表4参照）、事前学習済みモデルの学習目標の粒度の重要性が強調される。教師と生徒の特徴マップの位置合わせを確実にするため、各原画に同じ補強ビューを適用する。教師と生徒のモデル間で出力特徴マップの次元が異なることを可能にするために、生徒ネットワークの上に軽量プロジェクタを追加し、この方法をさらに一般化する。

生徒モデルのゴールは教師モデルを忠実に模倣することですが、生徒ネットワークをゼロからトレーニングすることで、異なる最適化経路が可能になります。この最適化経路の緩和により、教師ネットワークの表現力を最大限に維持しながら、生徒モデルがMIMと同様の特性を持つ可能性が得られる。

そして、より良い緩和を行い、望ましい帰納バイアスと正則化を導入することで、生徒モデルの移植性をさらに高めるために、以下の設計を提案する。

教師の特徴マップの標準化 事前学習されたモデルが異なると、特徴の大きさの次数が大きく異なる場合があり、ハイパーパラメータのチューニングが困難になる。さらに、小さな値にエンコードされた微妙な情報は、増幅されることなく生徒のネットワークにうまく抽出されないかもしれません。これらの問題を解決するために、ノンパラメトリックレイヤー正規化演算子[1]によって実装され、Tab. 7 (a).

蒸留では、生徒と教師の特徴マップの間に滑らかなℓ_1損失を用いる：

$$\mathcal{L}_{\text{distill}}(\mathbf{s},\mathbf{t}) = \begin{cases} \frac{1}{2}(g(\mathbf{s})-t')^2/\beta, & |g(\mathbf{s})-t'| \le \beta \\ (|g(\mathbf{s})-t'| - \frac{1}{2}\beta), & \text{otherwise} \end{cases}, \quad (1)$$

ここでβはデフォルトで2.0に設定され、t' = standardization (t)、sとtはそれぞれ生徒と教師ネットワークの出力特徴ベクトル、gはプロジェクタとして機能する1×1畳み込み層である。CLS]トークンはCLIPの事前学習時に大域的な画像情報を集約するため、[CLS]トークンの蒸留損失の重みを10.0倍に増幅する。

非対称なドロップパス率 特徴抽出フレームワークの2分岐構造は、教師と生徒のネットワークに非対称な正則化を可能にする。我々は、非対称ドロップパス[22]率の戦略を適用することで、より良い表現を学習できることを発見した。具体的には、表7 (b)に示すように、ViT-Bでは、生徒ブランチのドロップパス率を0.1とし、教師ブランチのドロップパス正則化を行わない戦略が最も有効である。7 (b).

相対位置バイアスの共有 元のCLIPモデル[42]は絶対位置符号化(APE)を採用しているが、最近の研究[38, 35, 34]では、相対位置バイアス(RPB)が下流タスクで利点を示すことを発見した。

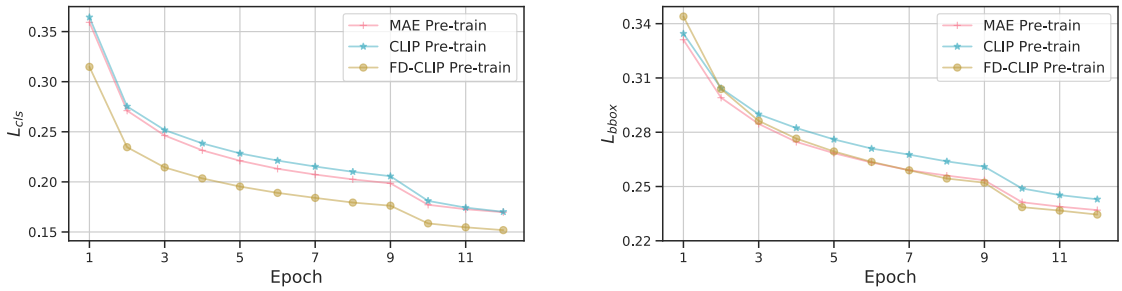


Figure 2: Fine-tuning MAE, CLIP and **FD**-CLIP on COCO object detection task. We visualize the loss curves of L_{cls} and L_{bbox} w.r.t the training epoch. Although CLIP pre-training is comparable to MAE pre-training on L_{cls}, it shows worse localization ability, reflected by the L_{bbox} curve.

ulate the input ratios (*i.e.* full image *vs*. partial image) and training target granularity (*i.e.* image-level *vs*. token-level) might be responsible for CLIP's inferior fine-tuning performance. While it is relatively easy to use partial inputs in CLIP, directly changing the training target from image-level to token-level would be challenging, as CLIP and existing vision-language training data are designed for image-level supervision. More importantly, re-training CLIP is costly that we would like to avoid.

Therefore, we propose to use distillation techniques, which are usually used for model compression, as a bridge for converting the training target granularity of CLIP models from image-level to token-level, while preserving the semantic information of the pre-trained model. To be specific, the pre-trained model serves as a frozen teacher, and a new same model with randomly initialized weights plays as the student, as illustrated in Fig. 1.

Instead of distilling logits like most previous distillation works [21, 51], we adopt the full output feature map of the pre-trained model as the distillation target, dubbed *"feature distillation"*. This approach allows us to work with any pre-trained model including those not having logits output. Moreover, distilling the feature map also leads to higher fine-tuning accuracy than only distilling a reduced single feature vector (see Tab. 4), emphasizing the importance of training target granularity for pre-trained models. To ensure that the feature maps of the teacher and student are aligned, we apply the same augmentation view to each original image. A light-weight projector is added on top of the student network to allow for different output feature map dimensions between the teacher and student models, further generalizing the method.

While the goal of the student model is to closely mimic the teacher model, training the student network from scratch allows a different optimization path. This relaxation of the optimization path provides the possibility for the student model to possess similar properties to those of MIM while maintaining the most of expressive power of the teacher network. Then we propose the following designs to make better relaxation and introduce desirable inductive biases and regularization, which further boosts the transferability of the student models.

Standardizing the teacher's feature map Different pre-trained models may have very different orders of feature magnitudes, which will make difficulties in hyper-parameter tuning. In addition, the subtle information that encoded in small values may not be well distilled into the student network without amplification. To solve these issues, we normalize the output feature map of the teacher network by a standardization operation, which is implemented by a non-parametric layer normalization operator [1] and proven to be important in Tab. 7 (a).

In distillation, we employ a smooth ℓ_1 loss between the student and teacher feature maps:

$$\mathcal{L}_{\text{distill}}(\mathbf{s}, \mathbf{t}) = \begin{cases} \frac{1}{2}(g(\mathbf{s}) - t')^2/\beta, & |g(\mathbf{s}) - t'| \leq \beta \\ (|g(\mathbf{s}) - t'| - \frac{1}{2}\beta), & \text{otherwise} \end{cases}, \quad (1)$$

where β is set 2.0 by default; $t' = \text{standardization}(\mathbf{t})$; $\mathbf{s}$ and $\mathbf{t}$ are output feature vectors of the student and teacher networks, respectively; g is a 1×1 convolution layer served as the projector. We amplify the distillation loss weight on `[CLS]` token by 10.0 as it aggregates the global image information during CLIP pre-training.

Asymmetric drop path rates The two-branch structure in the feature distillation framework allows for asymmetric regularization on the teacher and student networks. We find that applying a strategy of asymmetric drop path [22] rates can learn better representations. Specifically, the strategy of a drop path rate of 0.1 on the student branch with no drop path regularization on the teacher branch works best on ViT-B, as shown in Tab. 7 (b).

Shared relative position bias The original CLIP model [42] adopts the absolute position encoding (APE), but recent works [38, 35, 34] found the relative position bias (RPB) shows benefit on downstream tasks. Benefiting from the flexibility of feature distillation, we are able to

特徴抽出の柔軟性を利用して、我々は学生アーキテクチャにおける位置エンコーディング構成の影響を再検討することができる。特に、全レイヤーが同じ相対位置バイアス行列を共有する共有RPB構成を調査する。実験の結果、表7(c)に示すように、共有RPBが全体的に最も良い性能を発揮することが分かりました。7 (c). 我々は、共有RPBが、特に深いレイヤーのヘッドの多様性を高め（補足資料の図に示すように）、これが微調整性能の若干の向上に寄与している可能性が高いことを発見した。

4. 診断ツール

実験結果による特徴抽出の有効性の検証に加え、以下の診断ツールにより、学習された視覚表現のいくつかの興味深い性質を分析し、その背後にあるメカニズムの理解を提供する：

- 頭あたりの平均注意距離 [11]. この診断ツールは各パッチトークンが画像中で注目する平均相対距離を測定するもので、注目重みを用いて計算された各注目ヘッドの受容野サイズを部分的に反映する。CLS]トークンと各パッチ自体の測定は省略され、距離はピクセルレベルである。
- 各層の平均注意マップ[67]。各レイヤーの全ヘッドの平均アテンションマップを可視化する。注意マップには2つの共通パターンがある。対角線のパターンは、モデルが相対的な位置関係からの視覚的な手がかりにより多く依存していることを明らかにする。また、このパターンは、モデルの並進分散がより優れていることを示唆しており、これはしばしば、様々なダウンストリーム視覚タスクにとって有益な特性である。しかし、縦棒パターンは、固定された位置のパッチが他のすべての位置に強い影響を与えることを反映しており、これは並進変動である。対角線パターンでは、中央の対角線に集中することも、局所性事前分布を反映することができる、すなわち、中央に集中するほど局所性事前分布が強くなる。
- 正規化損失風景[33]。この診断ツールでは、学習されたモデルの重みは、程度の異なる一連のガウスノイズによって摂動される。33]に従い、各ノイズレベルは各フィルタのℓ_2ノルムに正規化され、異なるモデルの異なる重み振幅の影響を考慮する。視覚的に平坦な最小値は通常、より低いテスト誤差とより良い汎化能力に対応します[33]。

5. 実験と解析

本節では、CLIPの微調整性能が、特徴蒸留によるMIM法とのギャップを埋めることで改善できるかどうかを調べる。我々はまず、表4：蒸留対象の粒度に関するアブレーションが微調整性能に与える影響を調べる。入力比と訓練対象の粒度を蒸留する間に、我々の手法のいくつかの重要な設計を切除する。さらに、我々の診断ツールを用いて、蒸留前と蒸留後のモデルの詳細な分析を行う。

5.1. 実験設定

蒸留の設定。全ての実験において、1.28M ImageNet-1K トレーニング画像[10]を用いて特徴抽出を行う。すべての実験において100エポック（Tab. 1とTab. 2. デフォルトのモデルサイズはViT-B／16である。その他の詳細は補足資料を参照。

評価設定。ImageNet-1K分類[10]、ADE20Kセマンティックセグメンテーション[66]、 COCOオブジェクト検出とインスタンスセグメンテーション[36]、NYUv2深度推定[49]。

- ImageNet-1Kの分類。微調整のために、我々は[2]に従い、AdamWオプティマイザ[27]を使用し、学習率は層ごとに減衰させ、入力サイズは224×224とした。ViT-Bでは100エポック、ViT-Lでは50エポックで微調整する。線形プロービングについては、[17]に従い、LARSオプティマイザ[60]を使用し、基本学習率を0.1、ウェイト減衰を0とし、90エポック学習させた。トップ1の精度を報告する。その他の詳細は補足資料にある。
- ADE20Kセマンティックセグメンテーション。実験には[38]に従ってUPerNetフレームワーク[55]を使用する。AdamW[27]オプティマイザを採用し、学習長80K、バッチサイズ32、重み減衰0.05とする。その他のハイパーパラメータは、学習率4e-4、レイヤー減衰0.65、ドロップパス率0.2とした。学習時の入力画像サイズは512×512である。推論では[38]のシングルスケール検定に従う。検証セットでの平均IoUを報告する。
- COCO オブジェクト検出とインスタンス分割。我々は、MaskRCNNフレームワーク[19]を含む[6]のほとんどの設定に従う。高解像度のCOCO画像におけるグローバル自己注意によってもたらされるGPUメモリコストを削減するために、[38]のようなシフトウィンドウ注意を採用し、ウィンドウサイズを14と設定する。画像全体からの情報を集約するために、上部にグローバル自己注意層を追加する。検証セットにおけるBbox mAPとmask mAPを報告する。その他の詳細は補足資料にある。
- NYUv2深度推定。56, 25]の設定に従う。入力画像は480×480にランダムに切り出され、バッチサイズは24、最大学習率は5e-5、25エポック学習である。このタスクでRMSE(Root Mean Square Error)を評価する。その他の詳細は補足資料にある。

re-examine the impacts of position encoding configuration in the student architecture. In particular, we investigate a *shared RPB* configuration, where all layers share the same relative positional bias matrices. Our experiments show that the *shared RPB* performs best overall, as shown in Tab. 7 (c). We find that the *shared RPB* enhances the diversify of heads, particularly for the deeper layers (as shown in a figure in the supplementary material), which likely contributes to its slightly better fine-tuning performance.

4. Diagnostic tools

In addition to verifying the effectiveness of feature distillation through experimental results, we analyze several interesting properties of the learned visual representations to provide an understanding of the behind mechanism, by following diagnostic tools:

- *Average attention distances per head [11].* This diagnostic tool measures the average relative distance each patch token attends to in the image, which partially reflects the receptive field size for each attention head, computed using the attention weights. The `[CLS]` token and each patch itself are omitted in measurement, and the distances are pixel-level.
- *Average attention maps for each layer [67].* We visualize the attention maps averaged over all heads per layer. There are two common patterns in the attention maps: *diagonal* and *vertical-bar*. The *diagonal* pattern reveals that the model relies more on visual cues from relationships of relative locations. It also suggests better translation in-variance of the model, which is often a beneficial property for various down-stream visual tasks. However, the *vertical-bar* pattern reflects the strong impact of the patches in a fixed location to all other locations, which is translation variant. For the *diagonal* pattern, concentrating to a centered diagonal can also reflect *locality prior*, i.e. the more concentrated to the center, the stronger the locality prior.
- *Normalized loss landscapes [33].* In this diagnostic tool, the trained model weights are perturbed by a series of Gaussian noises with varying degrees. Following [33], each noise level is normalized to the ℓ_2 norm of each filter to account for the effects of varying weight amplitudes of different models. Visually flatter minimums usually correspond to lower test error and better generalization ability [33].

5. Experiments and Analysis

In this section, we investigate whether the fine-tuning performance of CLIP can be improved via bridging the gap with MIM methods through feature distillation. We firstly study the impacts on fine-tuning performance of different input ratios and training target granularity during distillation, and then we ablate several key designs in our method. Additionally, we provide a detailed analysis of the models before and after distillation, using our diagnosis tools.

5.1. Experimental Settings

Distillation settings. For all experiments, we perform feature distillation on 1.28M ImageNet-1K training images [10]. In ablation, we distill 100 epochs for all experiments, except for 300 epochs in Tab. 1 and Tab. 2. The default model size is ViT-B/16 if not mentioned else. Other details are in *supplemental materials*.

Evaluation settings. We include 4 evaluation benchmarks: ImageNet-1K classification [10], ADE20K semantic segmentation [66], COCO object detection and instance segmentation [36] and NYUv2 depth estimation [49].

- *ImageNet-1K classification*. For fine-tuning, we follow [2] to use the AdamW optimizer [27] with layer-wise decayed learning rates and an input size of 224×224. For ViT-B, we fine-tune it by 100 epochs, and for ViT-L, we fine-tune it by 50 epochs. For linear probing, we follow [17] to use the LARS optimizer [60] with a base learning rate of 0.1 and a weight decay of 0 training for 90 epochs. Top-1 accuracy is reported. Other details are in *supplemental materials*.
- *ADE20K semantic segmentation*. We follow [38] to use an UPerNet framework [55] for experiments. The AdamW [27] optimizer is employed with the training length of 80K, a batch size of 32, and a weight decay of 0.05. Other hyper-parameters are set as: learning rate 4e-4, layer decay 0.65, and drop path rate 0.2. In training, the input image size is 512×512. In inference, we follow the single-scale testing of [38]. Mean IoU on the validation set is reported.
- *COCO object detection and instance segmentation*. We follow the most settings in [6] including a Mask-RCNN framework [19] with $1\times$ schedule, multi-scale training and single-scale testing. To reduce the GPU memory cost brought by global self-attention on high-resolution COCO images, we adopt a shifted window attention like [38] and set the window size as 14. An additional global self-attention layer is added on the top to aggregate information from whole images. Bbox mAP and mask mAP on the validation set are reported. Other details are in *supplemental materials*.
- *NYUv2 depth estimation*. We follow the settings in [56, 25]. The input images are randomly cropped to 480×480 with a batch size of 24, maximal learning rate 5e-5 and 25-epoch training. We evaluate the RMSE (Root Mean Square Error) on this task. Other details are in *supplemental materials*.

モデルはImageNet-1Kデータセット[10]で100エポックで蒸留される。CLIPの微調整性能を高めるには、トークン・レベルのターゲットが不可欠である。

Method	IN-1K	ADE20K	COCO		NYUv2
	%	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
MAE [17]	83.6	48.1	46.5	40.9	0.383
[CLS] token	81.9	47.5	44.8	39.6	0.396
GAP feature	83.3	50.3	46.3	40.6	0.393
Full map	**84.4**	**51.8**	**47.9**	**42.2**	**0.350**

5.2. 学習対象の粒度と入力比率について

訓練対象の粒度 まず、訓練対象の粒度の影響を調査する。蒸留とターゲット粒度の影響を分離するために、3つの異なる蒸留ターゲットをアブレーションする：

- [CLS]トークン ビジュアルエンコーダの[CLS]トークンはCLIPにおいてユニークな役割を果たし、グローバルな画像情報を集約するだけでなく、豊富なセマンティクスを持つ言語モダリティにも対応する。この設定では、教師モデルからの[CLS]トークンの出力特徴量をターゲットとして、生徒モデルの対応する出力を導く。
- GAP特徴。この設定では、縮小された特徴ベクトルをターゲットとする。具体的には、グローバル平均プーリングレイヤーを特徴マップ全体に適用し、すべてのトークンからの情報を持つが、解像度が不足しているターゲットを構築する。
- フルマップ。これはFD-CLIPのデフォルト設定である。

表 4 に結果を示す。CLS]トークンをディスティリングすると、奥行き推定で改善が見られるが、他のタスクでは元のCLIPより悪い結果となった。比較すると、GAP特徴の蒸留はわずかな利点を示す。全特徴マップの使用は、すべての下流タスクにおいて、すべての蒸留ターゲットの中で最高の性能を示し、MAEモデルも上回る。

入力比率。マスクされた画像を抽出することにより、入力比率の影響をさらに調べる。表 5には、[75%, 50%, 25%, 0% (つまりFull)]の範囲でマスク比率を変化させた結果を示す。同じ学習エポックの下で、25%入力の設定が顕著に悪いことを除けば、フル画像と部分画像のディスティリングに大きな違いはないことがわかる。

以上の実験から、より良い微調整性能を達成するためには、学習ターゲットの粒度が重要であるという結論を得る。さらに、蒸留エポックを300に拡張し、最大のCLIPモデルであるViT-L／14で実験を行った。表1に示すように 表1に示すように、FD-CLIPはImageNet-1KとADE20Kにおいてオリジナル CLIPを2ポイント上回り、COCOでは約+3mAPの利得を得た。

表5：蒸留のための部分入力に対するアブレーション。モデルはImageNet-1Kデータセット[10]で100エポックで蒸留される。×x%マップは(100 − ×)%マスキングに等しい。†は教師モデルにもマスキングされた画像を入力することを意味し、そうでない場合はデフォルトで教師モデルにフル画像を使用する。

Method	IN-1K	ADE20K	COCO		NYUv2
	%	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
MAE [17]	83.6	48.1	46.5	40.9	0.383
25% input†	83.3	47.8	45.2	40.1	0.397
25% input	83.1	48.8	45.1	39.8	0.379
50% input	84.2	51.5	47.5	41.7	0.351
75% input	**84.4**	**52.0**	47.8	41.9	**0.347**
Full input	**84.4**	51.8	**47.9**	**42.2**	0.350

表6：ImageNet-1K [10]におけるCLIPの性能評価。特徴抽出により、CLIPの意味的能力はほぼ維持された。

Method	CLIP	**FD**-CLIP	Δ
Zero-shot (%)	68.6	68.0	-0.6
Linear probing (%)	79.5	80.1	+0.6

表7: 特徴蒸留における他の設計選択に関するアブレーション。太字はデフォルト設定。

(a) Normalization	None	ℓ_2 norm	Standardization
IN-1K (%)	83.5	83.9	**84.4**
(b) Std. / Tea. d.p.r	0.1 / 0.1	0.1 / 0	0.2 / 0
IN-1K (%)	84.0	**84.4**	84.0
(c) Position config.	APE	Non-shared RPB	Shared RPB
IN-1K (%)	84.0	83.9	**84.4**

また、NYUv2における深度推定のような低レベルタスクにおいても優位性を示し、MAEと比較してRMSEを0.033減少させた。ViT-Lモデルにスケールアップすると、ImageNet-1Kのファインチューニングで87.7%のトップ1精度を獲得し、オリジナルのCLIPを1.6%上回った（表2参照）。ImageNet-22K[10]での中間的なファインチューニングと、336×336へのより高いファインチューニング解像度を組み込むことで、ViT-LによるImageNet-1Kでの精度は89.0%に達する。ViT-L／14では、多重解像度FPNとモデルのパッチサイズ14との間に矛盾があり、2の指数乗にならないため、他の下流タスクは実施されていない。

全特徴マップの抽出はCLIPの事前学習目的とは異なるが、Tab. 6から、蒸留された学生モデルは、元のCLIPモデルのゼロショットと線形プロービング性能をほぼ維持していることがわかる。つまり、全特徴マップの蒸留は、MIM法の利点を取り入れつつ、CLIPモデルに取り込まれた多くの情報を継承することができ、それが優れた性能につながると考えられる。

Table 4: Ablation on distilling target granularity. The models are distilled on ImageNet-1K dataset [10] with 100 epochs. Token-level targets is vital to boost the fine-tuning performance of CLIP.

Method	IN-1K	ADE20K	COCO		NYUv2
	%	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
MAE [17]	83.6	48.1	46.5	40.9	0.383
`[CLS]` token	81.9	47.5	44.8	39.6	0.396
GAP feature	83.3	50.3	46.3	40.6	0.393
Full map	**84.4**	**51.8**	**47.9**	**42.2**	**0.350**

5.2. On training target granularity and input ratios

Training target granularity. We firstly investigate the impacts of training target granularity. To disentangle the impacts of distillation and target granularity, we ablate three different distillation targets:

- `[CLS]` *token*. The `[CLS]` token of the visual encoder plays a unique role in CLIP, which not only aggregates the global image information, but also aligns to the language modality with rich semantics. In this setting, we use the output feature of `[CLS]` token from the teacher model as the target to guide the corresponding output of the student model.
- *GAP feature*. In this setting, we use a reduced feature vector as the target. Specifically, a global average pooling layer is applied on the whole feature map to build targets with information from every tokens but lack of resolutions.
- *Full map*. We use the whole feature map without reduction as the target to create token-level supervision, which is the default setting in **FD**-CLIP.

Tab. 4 shows the results. Distilling `[CLS]` token shows an improvement on depth estimation, but performs worse on other tasks than original CLIP. In comparison, distilling GAP feature shows marginal benefits. The use of full feature map performs best among all distillation targets on all the downstream tasks, and also surpassing the MAE model.

Input Ratios. We further study the effects of input ratios by distilling masked images. Tab. 5 shows the results of different mask ratios on the student branch, ranging among [75%, 50%, 25%, 0% (*i.e.* Full)]. We find that under the same training epochs, there are no significant differences between distilling full images and partial images, except for the setting with only 25% input that is notably worse.

With the above experiments, we draw the conclusion that the training target granularity is crucial for achieving better fine-tuning performance. We further extend the distillation epochs to 300 and conduct experiments on the largest CLIP model, ViT-L/14. As shown in Tab. 1, **FD**-CLIP outperforms original CLIP by 2 points on ImageNet-1K and ADE20K and earns around +3 mAP gains on COCO. It also presents advantages on low-level tasks like depth estimation on NYUv2, reducing RMSE by 0.033 compared to MAE. When scaling up to ViT-L model, we earns 87.7% top-1 accuracy on ImageNet-1K fine-tuning, surpassing the original CLIP by 1.6% (see Tab. 2). Incorporating with intermediate fine-tuning on ImageNet-22K [10] and a higher fine-tuning resolution to 336 × 336, we reach 89.0% on ImageNet-1K with ViT-L. The other downstream tasks are not conducted on ViT-L/14 due to the inconsistency between the multi-resolution FPN and the model's patch size of 14, which is not an exponential power of 2.

Although distilling the full feature map is different from the pre-training objective of CLIP, Tab. 6 shows that the distilled student model has largely preserved zero-shot and linear probing performance of original CLIP model. That is, the distillation of the full feature map could inherit much information incorporated in the CLIP model while taking the advantages of MIM methods, which may lead to its superior performance.

Table 5: Ablation on partial inputs for distillation. The models are distilled on ImageNet-1K dataset [10] with 100 epochs. ×% map is equal to (100 − ×)% masking. † means we also input a masked image into the teacher model, otherwise we use the full image for the teacher model by default.

Method	IN-1K	ADE20K	COCO		NYUv2
	%	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
MAE [17]	83.6	48.1	46.5	40.9	0.383
25% input†	83.3	47.8	45.2	40.1	0.397
25% input	83.1	48.8	45.1	39.8	0.379
50% input	84.2	51.5	47.5	41.7	0.351
75% input	**84.4**	**52.0**	47.8	41.9	**0.347**
Full input	**84.4**	51.8	**47.9**	**42.2**	0.350

Table 6: Evaluating the distilled CLIP performance on ImageNet-1K [10]. The feature distillation could largely preserve the semantic capability of CLIP.

Method	CLIP	**FD**-CLIP	Δ
Zero-shot (%)	68.6	68.0	-0.6
Linear probing (%)	79.5	80.1	+0.6

Table 7: Ablation on other design choices in feature distillation. **Bold** ones are our default settings.

(a) Normalization	None	ℓ_2 norm	Standardization
IN-1K (%)	83.5	83.9	**84.4**
(b) Std. / Tea. d.p.r	0.1 / 0.1	0.1 / 0	0.2 / 0
IN-1K (%)	84.0	**84.4**	84.0
(c) Position config.	APE	Non-shared RPB	Shared RPB
IN-1K (%)	84.0	83.9	**84.4**

表8：DINO[3]とDeiT[51]で事前学習したViT-Bに特徴蒸留を適用。表8：DINO[3]とDeiT[51]で事前学習されたViT-Bの特徴抽出。

Method	IN-1K	ADE20K	COCO		NYUv2
	%	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
DINO [3]	82.8	46.2	45.8	40.7	0.412
FD-DINO	83.8	47.7	46.1	40.9	0.394
Δ	**↑1.0**	**↑1.5**	**↑0.3**	**↑0.2**	**↓0.018**
DeiT [51]	81.8	47.0	45.8	40.7	0.403
FD-DeiT	83.0	48.0	46.4	41.0	0.404
Δ	**↑1.2**	**↑1.0**	**↑0.6**	**↑0.3**	**↑0.001**

5.3. デザイン選択肢の削除

この節では、特徴蒸留フレームワークの他の設計について述べる。全ての実験は、ViT-Bと100エポック学習を用いたImageNet-1Kデータセットで行われる。
教師特徴の正規化について 表 7（a）は、教師特徴マップの正規化を行うかどうかの効果を示している。教師特徴マップの正規化により、元の特徴マップを使用する場合よりも+0.9%の改善が得られる。ℓ_2ノルムと標準化の2つの正規化アプローチを比較すると、後者の方が+0.5%の改善が見られる。また、正規化により、特徴蒸留のハイパーパラメータは事前学習モデルの影響を受けなくなる。

非対称ドロップパス率について。表 7（b)は、ドロップパスの正則化の程度による効果を示している。生徒ネットワークにドロップパスの正則化を適度に加えることは、おそらくオーバーフィッティングの緩和により、有益であろう。しかし、教師モデルにドロップパスの正則化を加えると性能が低下することから、正確な教師信号が有益であることがわかる。したがって、我々はデフォルトでこの非対称ドロップパス率戦略を採用する。

位置エンコーディングの構成について 表 7（c)は、生徒ネットワークにおける位置エンコーディングの構成を変化させた場合の効果を示している。その結果、共有相対位置バイアス（共有RPB）構成が他を凌駕することが明らかになった。それにもかかわらず、すべての構成が非常に良好に動作するため、適切な位置エンコーディング構成は、特徴蒸留の成功の決定的な要因ではない。

5.4. より多くのモデルでの評価

Tab.1とTab. 1とTab. 2に示す実験から、CLIPモデルにおける特徴抽出の有効性が明らかになった。本研究の動機はCLIPの微調整性能を向上させることであるが、特徴蒸留アプローチは他の事前学習モデルでも有効である。Table. 8では、DINO [3]とDeiT [51]に特徴蒸留を適用し、様々な下流タスクで一貫した改善を観察している。これは、特徴抽出アプローチが、異なる事前学習対象で事前学習されたモデルに対しても有効であることを示している。また、MAE [17]についても同様の実験を行っている、

表9：特徴抽出により、SwinV2-Gモデル[37]も様々なタスクで改善される。

Method	IN-1K	COCO		ADE20K
		AP_{box}	AP_{mask}	mIoU
GLIPv2-CoSwin-H [65]	-	62.4	-	-
Florence-CoSwin-H [61]	-	62.4	-	-
DINO-Swin-L [64]	-	63.3	-	-
MaskDINO-Swin-L [32]	-	-	54.7	60.8
ViT-Adapter-L [8]	-	-	-	60.5
SwinV2-G [37]	89.2	63.1	54.4	59.9
FD-SwinV2-G	**89.4** (+0.2)	**64.2** (+1.1)	**55.4** (+1.0)	**61.4** (+1.5)

を付録に示す。FD-MAEはほとんどのタスクでわずかな利得を獲得し、我々の手法の利得は主にトークン・レベルのタスクから得られるという我々の観測を検証する。
また、我々は30億パラメータのSwinV2-Gを改良し、ADE20Kセマンティックセグメンテーションと COCOオブジェクト検出で61.4mIoUと64.2mAPを達成し（オリジナルのSwin V2[37]と同じUperNet ／ HTC++フレームワークと同じ評価設定を使用）、Tab.9に示すように、それぞれ（マスク）DINO[64, 32]で報告された以前の最先端よりも+0.6mIoUと+0.9mAP高い新レコードを作成した。9. これらの結果は、異なる事前学習法とモデルアーキテクチャに対する我々のアプローチの一般的な適用可能性を示唆している。

5.5. Analysis

広範な実験結果から、特徴抽出はCLIPモデルの微調整性能を促進できることが示された。本節では、特徴蒸留がモデルの挙動にどのような影響を与えるかを理解するために、第4節で述べたツールを用いてモデルを診断する。全ての解析は50,000枚のImageNet-1K検証画像に対して行われた。

多様化する注意の頭部 我々はまず、ネットワーク層に対する異なるヘッドの注意の多様性を調べる。図3は、MAE、CLIP、特徴抽出CLIP（FD-CLIP）それぞれの異なるネットワーク層におけるヘッドごとの平均注意距離を示している。浅い層では、どのモデルの学習表現も頭部間で乖離している。しかし、CLIPモデルの深い層では、異なる頭部の出席距離に関する多様性が急速に収束している。直感的には、収束した表現は、モデルの容量が十分に利用されておらず、冗長性がある可能性を示している[56]。これと比較すると、FD-CLIPはこの問題を軽減し、その表現はMAEの表現に近い。

並進不変性の強化 図4は、異なるモデルの平均アテンションマップを示す。CLIPモデルと比較して、MAE事前学習モデルは、相対的な位置からの視覚的な手がかりをより重視し、他よりも多くの対角線パターンを示すことから、MAEの並進不変性がより優れていることが示唆される。このローカリティ特性は、きめ細かな定位能力を必要とする下流のタスクに役立つ可能性がある。

Table 8: Applying feature distillation on ViT-B pre-trained with DINO [3] and DeiT [51]. The models are distilled on ImageNet-1K dataset [10] with 300 epochs.

Method	IN-1K	ADE20K	COCO		NYUv2
	%	mIoU	AP_{box}	AP_{mask}	RMSE (↓)
DINO [3]	82.8	46.2	45.8	40.7	0.412
FD-DINO	83.8	47.7	46.1	40.9	0.394
Δ	↑**1.0**	↑**1.5**	↑**0.3**	↑**0.2**	↓**0.018**
DeiT [51]	81.8	47.0	45.8	40.7	0.403
FD-DeiT	83.0	48.0	46.4	41.0	0.404
Δ	↑**1.2**	↑**1.0**	↑**0.6**	↑**0.3**	↑**0.001**

Table 9: Feature distillation also improves the advanced SwinV2-G model [37] on various downstream tasks.

Method	IN-1K	COCO		ADE20K
		AP_{box}	AP_{mask}	mIoU
GLIPv2-CoSwin-H [65]	-	62.4	-	-
Florence-CoSwin-H [61]	-	62.4	-	-
DINO-Swin-L [64]	-	63.3	-	-
MaskDINO-Swin-L [32]	-	-	54.7	60.8
ViT-Adapter-L [8]	-	-	-	60.5
SwinV2-G [37]	89.2	63.1	54.4	59.9
FD-SwinV2-G	**89.4** (+0.2)	**64.2** (+1.1)	**55.4** (+1.0)	**61.4** (+1.5)

5.3. Ablation of design choices

In this section, we ablate other designs of in our feature distillation framework. All experiments are performed on ImageNet-1K dataset using ViT-B and 100-epoch training.

On the normalization of teacher features. Tab. 7 (a) ablates the effect of whether and how to perform teacher feature map normalization. Teacher feature map standardization brings +0.9% improvement over using the original feature maps. Comparing two normalization approaches of ℓ_2 norm and standardization, the latter one shows a gain of +0.5%. Normalization also makes feature distillation hyper-parameters insensitive to the pre-training models.

On asymmetric drop path rates. Tab. 7 (b) ablates the effect of different degrees of drop path regularization. Moderately adding the drop path regularization on the student network would be beneficial, possibly due to the relief of over-fitting. However, adding drop path regularization on the teacher model damages the performance, indicating that an accurate teacher signal is beneficial. Therefore, we adopt this asymmetric drop path rate strategy by default.

On position encoding configurations. Tab. 7 (c) ablates the effect of varying position encoding configurations in the student network. The results reveal that the shared relative position bias (*shared RPB*) configuration outperforms others. Nonetheless, all configurations perform quite well, so the proper position encoding configuration is not the decisive factor for the success of feature distillation.

5.4. Evaluation on more models

Experiments shown in Tab. 1 and Tab. 2 reveal the effectiveness of feature distillation on CLIP models. While the motivation of this work is to improve the fine-tuning performance of CLIP, the feature distillation approach also works with other pre-training models. In Tab. 8, we apply the feature distillation on DINO [3] and DeiT [51] and observe consistent improvements on various downstream tasks. It shows that the feature distillation approach is also effective on models pre-trained with different pre-training objects. We also conduct similar experiments on MAE [17], shown in appendix. **FD**-MAE earns marginal gain on most tasks, verifying our observations that the gain of our method is largely from a token-level task.

We also improve the 3-billion-parameter SwinV2-G to achieve **61.4 mIoU** and **64.2 mAP** on ADE20K semantic segmentation and COCO object detection (using the same UperNet / HTC++ framework and the same evaluation settings as the original Swin V2 [37]), creating new records with +0.6 mIoU and +0.9 mAP higher than previous state-of-the-art reported in (Mask) DINO [64, 32], respectively, as shown in Tab. 9. These results suggest the general applicability of our approach to different pre-training methods and model architectures.

5.5. Analysis

Extensive experimental results have shown that feature distillation can facilitate the fine-tuning performance of CLIP models. In this section, we diagnose the models with tools mentioned in Sec. 4 to understand how feature distillation affects the model behaviors. All the analyses are performed on 50,000 ImageNet-1K validation images.

Diversified attention heads. We firstly examine the attention diversity of different heads w.r.t. network layers. Fig. 3 shows the average attention distance per head in different network layers of MAE, CLIP and feature distilled CLIP (**FD**-CLIP), respectively. At shallow layers, the learned representations of all models are diverged across heads. However, the diversity on attended distances of different heads is rapidly converging in deep layers of the CLIP model. Intuitively, the converged representation indicating the model capacity is not fully utilized and may have redundancy [56]. In comparison, **FD**-CLIP alleviates this issue and its representations are more similar to the ones in MAE.

Enhanced translational invariance. Fig. 4 shows the average attention maps of different models. Compared to CLIP models, the MAE pre-trained model focuses more on the visual cues from the relative locations, and shows more *diagonal* patterns than others, suggests better translational invariance of MAE. This locality property may benefit down-

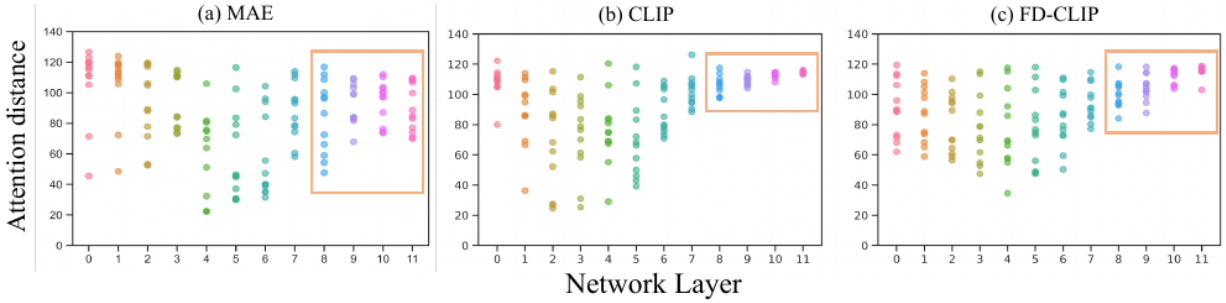


図3：(a)MAE[17]、(b)CLIP[42]、(c)FD-CLIPの各レイヤー深度におけるヘッドあたりの平均注意距離。距離はピクセル・レベルで測定されている。

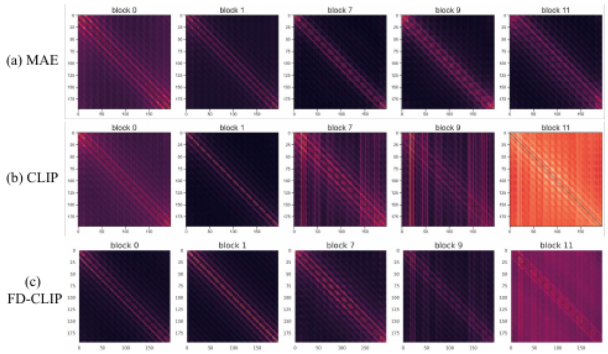


図4：(a)MAE [17]、(b)CLIP [42]、(c)FD-CLIPの平均注意マップ。マップは全頭、全画像の平均である。第0、1、7、9、11層の代表的な5層が、スペースを節約するために選択されている。完全なアテンション・マップは補足資料にある。

一方、CLIPの注意マップでは、より深い層（例えばブロック7 11）に縦棒パターンが多く、これはCLIPの特徴が絶対位置の特定のパッチに支配されていることを示している。これは、CLIPの特徴量が、絶対的な位置にある特定のパッチを支配していることを示している。縦棒パターンは、特徴量の蒸留後、部分的に消失し、蒸留されたモデルは、MAEが行っているような相対的な位置からの視覚的手がかりの符号化関係にもっと依存し、より優れた並進不変性を示すことが明らかになった。

平坦化された損失風景 図5は、異なるモデルの損失と精度のランドスケープ[33]を視覚化したものである。その結果、MAEとFD-CLIPは、元のモデルよりも比較的平坦であることがわかった。この観察結果は、より良い微調整精度とも一致する。

6. Conclusion

本論文では、CLIPモデルに古典的な特徴蒸留のフレームワークを採用することで、CLIPモデルの微調整性能を向上させると同時に、CLIPモデル本来のse-se-se-se-se-を継承することを目指す。

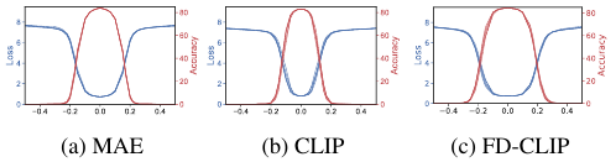


図5：(a)MAE [17]、(b)CLIP [42]、(c)FD-CLIPの損失／精度のランドスケープ [33]。各プロットには、ランダムに生成された5つの方向を使用した5つの風景がある。

マンティックの能力 CLIP手法とMIM手法の分類・局在化タスクにおける成分の違いと動作の違いを分析することで、トークンレベルのターゲット粒度を持つタスクが、MIM手法の成功の鍵の1つであり、特にその素晴らしい微調整性能の鍵であることがわかった。この観点から、我々は、事前に学習されたCLIPモデルにトークンレベルのタスクを提供するために、いくつかの重要な設計を持つ古典的な特徴蒸留フレームワークを導入した。その結果、オリジナルのCLIPモデルと比較して、様々な下流タスクにおいて一貫した明確な改善を得ることができ、ImageNet1K分類におけるゼロショットや線形プロービングのように、その意味的能力はほぼ維持された。さらに、いくつかの注意と最適化に関連する診断ツールを用いて、FD-CLIPとMIMおよびCLIPを分析した。その結果、FD-CLIPはMIMとより類似したパターンを共有することが明らかになった。さらに、DeiT、DINO、先進的なSwinV2-Gを含む、より多様なモデルにフレームワークを一般化し、一貫した利点を観察した。

限界 特徴抽出後の性能向上は顕著であるが、モデル学習パイプラインはより複雑になり、追加学習コストが必要となる。このコストについては付録を参照されたい。また、MAEやFDCLIPを分析する診断ツールも直感的なものであり、微調整の性能を直接示すことはできないかもしれない。

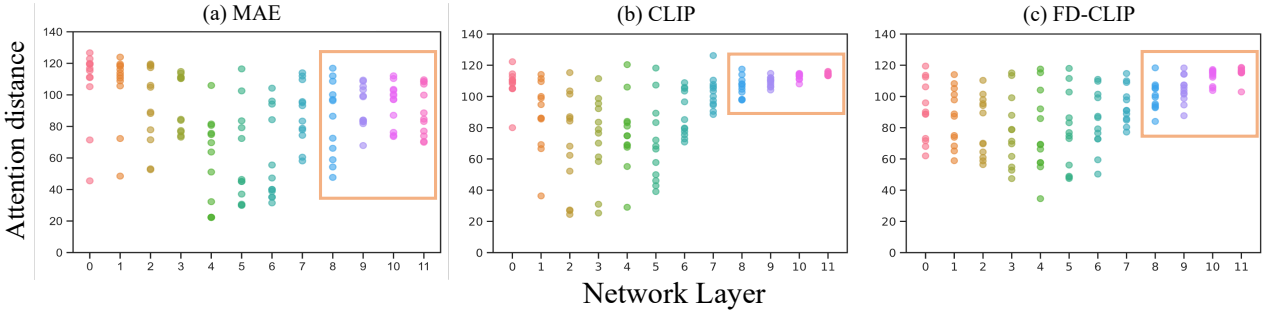


Figure 3: The average attention distance per head at each layer depth on (a) MAE [17], (b) CLIP [42] and (c) **FD**-CLIP. The distances are measured on the pixel level.

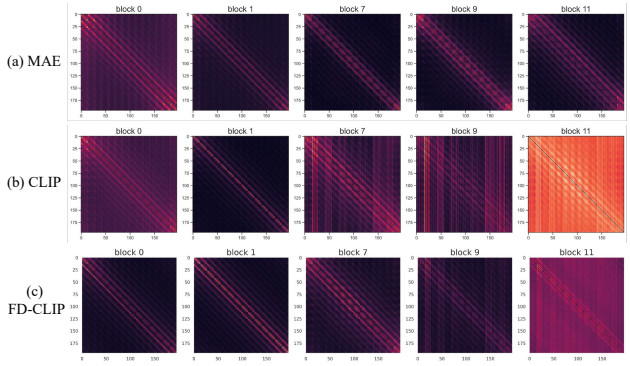
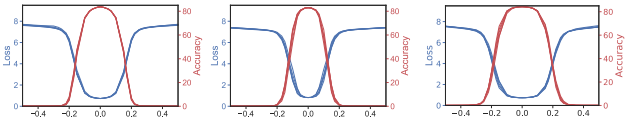


Figure 4: The average attention maps on (a) MAE [17], (b) CLIP [42] and (c) **FD**-CLIP. The maps are averaged over all heads and all images. Five representative layers, 0th, 1st, 7th, 9th, 11th, are selected to save the space. Full attention maps can be found in the *supplementary materials*.



(a) MAE (b) CLIP (c) FD-CLIP

Figure 5: The loss / accuracy landscapes [33] of (a) MAE [17], (b) CLIP [42] and (c) **FD**-CLIP, where x-axis represents the noise strength and y-axis is the loss / accuracy. Each plot has 5 landscapes using 5 randomly generated directions.

stream tasks that requires a fine-grained localization ability. In contrast, the attention maps of CLIP have much more *vertical-bar* patterns in deeper layers (*e.g.* block 7-11), which indicates the CLIP features are dominated by certain patches on absolute locations. The *vertical-bar* patterns partly disappear after feature distillation, revealing that the distilled model relies more on encoding relationship of visual cues from relative locations like what MAE does, and shows better translational invariance.

Flattened loss landscapes. Fig. 5 visualizes the loss and accuracy landscapes [33] of different models. It turns out that the landscapes of MAE and **FD**-CLIP are relatively flatter than original ones, which generally reflects its optimization friendliness and better generalization. This observation is also consistent with their better fine-tuning accuracy.

6. Conclusion

This paper seeks to adopt a classical feature distillation framework on CLIP models to improve their fine-tuning performance and simultaneously inherit the original semantic capability. By analyzing the ingredient differences and behaviors differences on classification and localization tasks between CLIP and MIM methods, we found that a task with token-level target granularity is one of the key to the success of MIM methods, especially to their impressive fine-tuning performance. From this perspective, we introduced the classical feature distillation framework with several crucial designs to provide a token-level task for pre-trained CLIP models. We gained consistent and clear improvements on various downstream tasks compared to the original CLIP models and largely preserved their semantic capability, like zero-shot and linear probing on ImageNet-1K classification. Besides, we analyzed **FD**-CLIP with MIM and CLIP using several attention- and optimization-related diagnosing tools. The visualizations revealed that after distillation, **FD**-CLIP shares more similar patterns with MIM. Moreover, we further generalized the framework to more various models including DeiT, DINO and the advanced SwinV2-G and observed consistent gains.

Limitations Although the performance improvements are noticeable after feature distillation, the model training pipeline becomes more complicated and additional training cost is required, *e.g.* 3% more compared to CLIP pre-training. We further discuss the cost in the appendix. Besides, the diagnosing tools which analyze the MAE and **FD**-CLIP are also intuitive and may not be able to indicate the fine-tuning performance directly.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 3, 5

[3] 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. 2, 3, 7

[4] (1)自己教師付きモデルは強い半教師付き学習者である。Riemannian walk for incremental learning： 忘却とintransigenceを理解する。Proceedings of the European Conference on Computer Vision (ECCV), pages 532-547, 2018. 3

[5] (1)画像は16x16語の価値がある: スケールの大きな画像認識のための変換器。このような場合、「自己監視モデル」と「半教師付き学習モデル」を比較することで、「自己監視モデル」と「半教師付き学習モデル」を比較することができる。(注1) 本論文は、本論文の一部である。3

[6] このような視覚的な共通感覚は、視覚的な共通感覚を学習するための、視覚的な共通感覚を評価するための、視覚的な共通感覚を学習するためのビデオデータベースである。(訳注:この論文では、"text autoencoder for self-supervised representation learning "と表記している。) arXiv preprint arXiv:2202.03026, 2022. 5

[7] Xu Chen, Yongfeng Zhang, Hongteng Xu, Zheng Qin, and Hongyuan Zha. (1)嗜好の変化に対する嗜好の変化,(2)嗜好の変化に対する嗜好の変化. ACM Transactions on Information Systems (TOIS), 37(1):1-28, 2018. 3

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 7

[9] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1, 2, 3, 5, 6, 7

[11] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit、Neil Houlsby。画像は16x16語の価値がある： 画像は16x16語の価値がある: Transformers for image recognition at scale. 国際学習表現会議, 2021. 2, 3, 5

[12] このような、「自己教師付き表現学習」のためのコンテキスト・オートエンコーダは、「自己教師付き表現学習」のためのコンテキスト・オートエンコーダである。このような場合、このような閾値は、閾値を超える閾値となる。In International Conference on Machine Learning, pages 1607-1616. PMLR, 2018. 2, 3

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013. 1

[14] Gabriel Goh、Nick Cammarata、Chelsea Voss、Shan Carter、Michael Petrov、Ludwig Schubert、Alec Radford、Chris オラ 人工ニューラルネットワークにおけるマルチモーダルニューロン. Distill, 6(3):e30, 2021. 3

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. Proceedings of the IEEE international conference on computer vision, pages 5842-5850, 2017. 3

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 1, 3, 5, 6, 7, 8

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5

[20] Geoffrey Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006. 1

[21] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 2, 3, 4

[22] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 4

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. 2021. 3

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[25] Doyeon Kim, Woonghyun Ga, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 5

[26] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 3

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 3

[29] Simon Kornblith, Jonathon Shlens, and Quoc V Le. より良いイメージネットモデルはより良く転送するか？In Proceedings of

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2, 3, 5

[3] 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. 2, 3, 7

[4] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–547, 2018. 3

[5] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 3

[6] 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. 5

[7] Xu Chen, Yongfeng Zhang, Hongteng Xu, Zheng Qin, and Hongyuan Zha. Adversarial distillation for efficient recommendation with external knowledge. *ACM Transactions on Information Systems (TOIS)*, 37(1):1–28, 2018. 3

[8] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022. 7

[9] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1, 2, 3, 5, 6, 7

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2, 3, 5

[12] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018. 2, 3

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013. 1

[14] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 3

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3

[16] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 3

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 1, 3, 5, 6, 7, 8

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *CVPR*, 2020. 3

[19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 5

[20] Geoffrey Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006. 1

[21] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. 2, 3, 4

[22] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 4

[23] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. 2021. 3

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[25] Doyeon Kim, Woonghyun Ga, Pyungwhan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *arXiv preprint arXiv:2201.07436*, 2022. 5

[26] Jangho Kim, SeongUk Park, and Nojun Kwak. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems*, 31, 2018. 3

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[28] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019. 3

[29] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of*

the IEEE／CVF conference on computer vision and pattern recognition, pages 2661-2671, 2019. 3

[30] アレックス・クリシェフスキー（Alex Krizhevsky）、イリヤ・スーツケバー（Ilya Sutskever）、ジェフリー・E・ヒントン（Geoffrey E Hinton）。を用いた画像分類。このような場合、「曖昧さ」が「曖昧さ」である可能性が高い。2012. 1, 3

[31] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *International Conference on Learning Representations (ICLR)*, 2022. 3

[32] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022. 7

[33] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets, 2017. 3, 5, 8

[34] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ArXiv*, abs/2203.16527, 2022. 4

[35] 李陽昊、呉超遠、范浩基、Karttikeya Mangalam、Bo Xiong、Jitendra Malik、Christoph Feichtenhofer。Mvitv2：分類と検出のための改良されたマルチスケールビジョン変換器。コンピュータビジョンとパターン認識に関するIEEE／CVF会議予稿集、4804-4814ページ、2022年。4

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C L awrence Zitnick. Microsoft coco：Microsoft coco：Common objects in context. In European conference on computer vision, pages 740-755. Springer, 2014. 2, 3, 5

[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 2, 3, 7

[38] 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. pages 10012–10022, 2021. 4, 5

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[40] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3

[41] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3, 4, 8

[43] このような場合、視覚変換は畳み込みのように見えるのだろうか？視覚変換器は畳み込みニューラルネットワークのように見えるか？アドヴァンス

in Neural Information Processing Systems, 34:12116–12128, 2021. 3

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[46] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3

[47] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. 3

[48] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 3

[49] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 2, 5

[50] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. arXiv preprint arXiv:2012.12877, 2020. 2, 3, 4, 7

[52] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 2

[53] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 3

[54] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2021. 2

[55] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 5

[56] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022. 3, 5, 7

[57] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level

the IEEE/CVF conference on computer vision and pattern recognition, pages 2661–2671, 2019. 3

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105. 2012. 1, 3

[31] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *International Conference on Learning Representations (ICLR)*, 2022. 3

[32] Feng Li, Hao Zhang, Huaizhe xu, Shilong Liu, Lei Zhang, Lionel M. Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation, 2022. 7

[33] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets, 2017. 3, 5, 8

[34] Yanghao Li, Hanzi Mao, Ross B. Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *ArXiv*, abs/2203.16527, 2022. 4

[35] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 4

[36] 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, 3, 5

[37] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 2, 3, 7

[38] 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. pages 10012–10022, 2021. 4, 5

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[40] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 3

[41] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 2, 3, 4, 8

[43] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in Neural Information Processing Systems*, 34:12116–12128, 2021. 3

[44] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2

[46] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 3

[47] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can CLIP benefit vision-and-language tasks? In *International Conference on Learning Representations*, 2022. 3

[48] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017. 3

[49] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. *ECCV (5)*, 7576:746–760, 2012. 2, 5

[50] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3

[51] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020. 2, 3, 4, 7

[52] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 2

[53] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 3

[54] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models, 2021. 2

[55] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418–434, 2018. 5

[56] Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, and Yue Cao. Revealing the dark secrets of masked image modeling. *arXiv preprint arXiv:2205.13543*, 2022. 3, 5, 7

[57] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level

教師なし視覚表現学習のための一貫性。このような画像生成のための学習は、画像生成のための学習と、画像生成のための学習と、画像生成のための学習と、画像生成のための学習とがある。 3

[58] 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 *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[59] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 3

[60] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 5

[61] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 7

[62] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan： 条件付き画像生成のための継続学習。Proceedings of the IEEE／CVF International Conference on Computer Vision, pages 2759-2768, 2019. 3

[63] 蔡暁華、Alexander Kolesnikov、Neil Houlsby、Lucas Beyer。ビジョン変換器のスケーリング。このような、視覚変換器のスケーリングは、コンピュータビジョンとパターン認識に関するIEEE／CVF会議予稿集、12104-12113ページ、2022年。2

[64] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 7

[65] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding, 2022. 7

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 2, 3, 5

[67] 楊林杰、梁暁晨、侯奇彬、馮佳志。Deepvit： arXiv preprint arXiv:2103.11886, 2021. 3, 5

consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021. 3

[58] 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 *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 3

[59] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 3

[60] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017. 5

[61] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 7

[62] Mengyao Zhai, Lei Chen, Frederick Tung, Jiawei He, Megha Nawhal, and Greg Mori. Lifelong gan: Continual learning for conditional image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2759–2768, 2019. 3

[63] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 2

[64] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022. 7

[65] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding, 2022. 7

[66] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal on Computer Vision*, 2018. 2, 3, 5

[67] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 3, 5