AI Dev. Group CV team

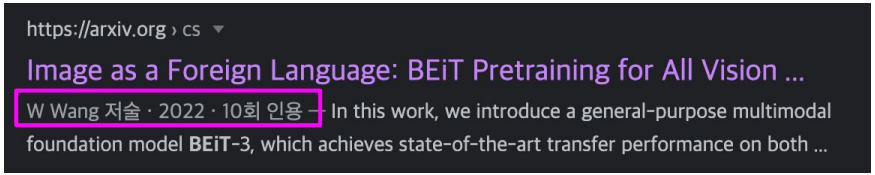
- Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks
- Vision Transformer Adapter for Dense Predictions

2022-11-16 | 이지현

Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks

Wenhui Wang[*], **Hangbo Bao**[*], **Li Dong**[*], **Johan Bjorck**, **Zhiliang Peng**, **Qiang Liu**
Kriti Aggarwal, **Owais Khan Mohammed**, **Saksham Singhal**, **Subhojit Som**, **Furu Wei**[†]
Microsoft Corporation
https://aka.ms/beit-3

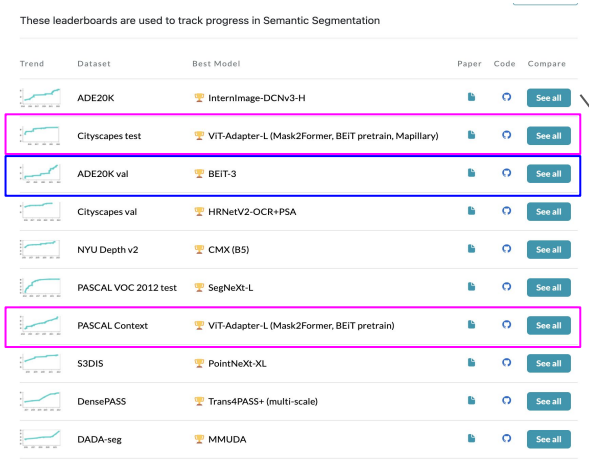
https://arxiv.org/abs/2208.10442
https://github.com/microsoft/unilm/tree/master/beit

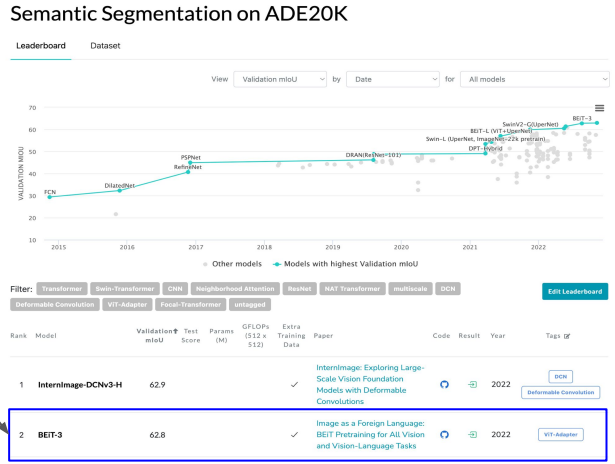


Content

Contents

- Image as a Foreign Language: BEIT Pretraining for All Vision and Vision-Language Tasks
  - 2022 ViT 기반 최신 모델
  - ADE20K val data 1위, ADE20K 2위 (최근에 SoTA 바뀜)
  - official git code (O)
- Vision Transformer Adapter for Dense Predictions
  - 2022 ViT 기반 최신 모델
  - PASCAL Context 1위, 2위, Cityscapes test 1위
  - official git code (O)





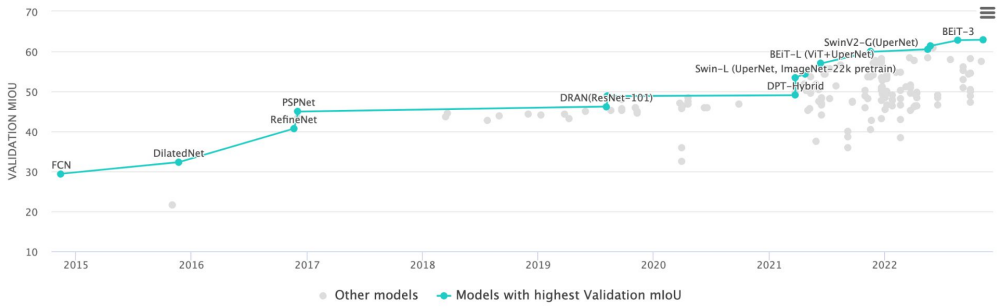
Introduction

- **General-purpose multimodal foundation model BEiT-3**
- (ADE20K) 짧은 사이에 SoTA 가 바뀌어버린 ..

Semantic Segmentation on ADE20K

Leaderboard Dataset

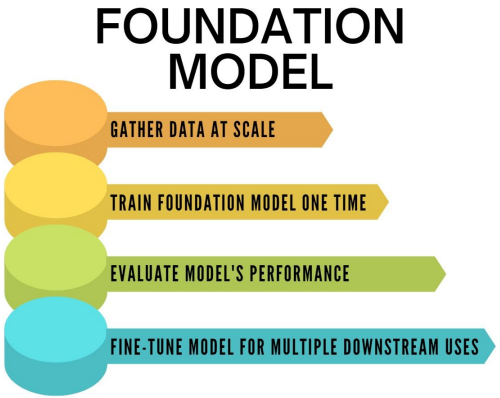
View Validation mIoU by Date for All models



1	InternImage-DCNv3-H	62.9	✓	InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions		2022	DCN, Deformable Convolution
2	BEiT-3	62.8	✓	Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks		2022	ViT-Adapter
3	FD-SwinV2-G	61.4	✓	Contrastive Learning Rivals Masked Image Modeling in Fine-tuning via Feature Distillation		2022	Swin-Transformer

Introduction

- **General-purpose multimodal foundation model BEiT-3**
- Foundation model
  - The future is models that are trained on a broad set of unlabeled data that can be used for different tasks, with minimal fine-tuning. These are called foundation models, a term first popularized by the Stanford Institute for Human-Centered Artificial Intelligence.
  - 말 그대로 AI 모델의 많은 applications 에 foundation 이 될 수 있음
  - 최소한의 fine-tuning 으로 다양한 downstream task 에 적용할 수 있도록, unlabeled large-scale 데이터셋으로 학습시킨 모델 (일반적으로 self-supervised learning)
  - 초기 foundation model 로는 BERT, GPT-3 등이 있음. 최근 multimodal foundation model 로 DALL-E, Flamingo, Florence 등이 있음



Introduction

- Semantic Segmentation

Benchmarks

Add a Result

These leaderboards are used to track progress in Semantic Segmentation

Trend	Dataset	Best Model	Paper	Code	Compare
	ADE20K	InternImage-DCNv3-H			See all
	Cityscapes test	ViT-Adapter-L (Mask2Former, BEiT pretrain, Mapillary)			See all
	ADE20K val	BEiT-3			See all
	Cityscapes val	HRNetV2-OCR+PSA			See all
	NYU Depth v2	CMX (B5)			See all
	PASCAL VOC 2012 test	SegNeXt-L			See all
	PASCAL Context	ViT-Adapter-L (Mask2Former, BEiT pretrain)			See all
	S3DIS	ArgosNet			See all
	DensePASS	Trans4PASS+ (multi-scale)			See all
	DADA-seg	MMUDA			See all

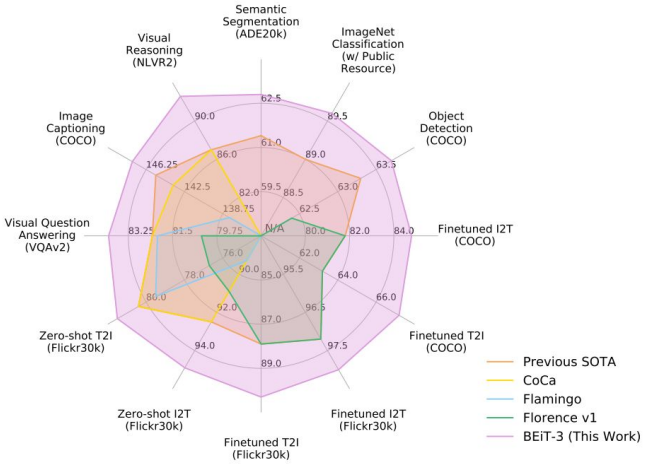


Figure 1: BEIT-3 achieves state-of-the-art performance on a broad range of tasks compared with other customized or foundation models. I2T/T2I is short for image-to-text/text-to-image retrieval.

BEiT-3: A General-Purpose Multimodal Foundation Model

- **Overview**
  - **BEiT-3 는 monomodal (image, text) 와 multimodal (image-text pair) 에 대해 masked data modeling 을 통해 pre-trained**
    - Shared Multiway Transformer network 사용
  - **학습된 모델은 다양한 vision, vision-language downstream tasks 로 transfer 가능**

- **Advance the big convergence from three aspects**
  - Backbone architecture, pretraining task, and model scaling up

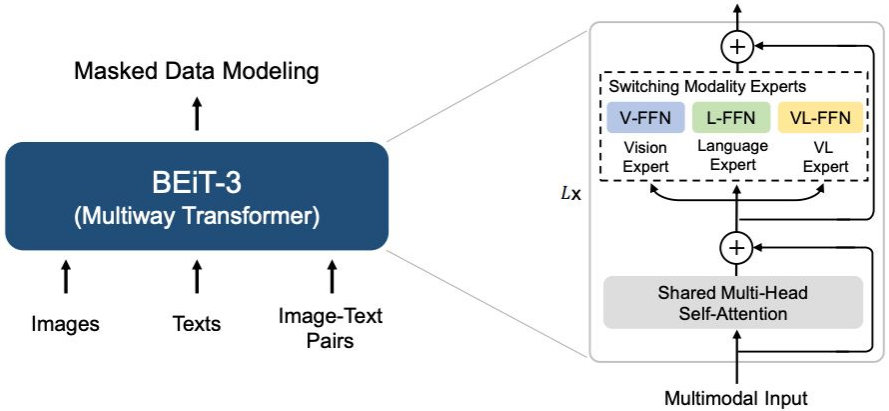


Figure 2: Overview of BEIT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.

BEiT-3: A General-Purpose Multimodal Foundation Model

- **Advance the big convergence from three aspects: 1) <u>Backbone architecture: Multiway Transformers</u>**
  - 다양한 모달리티를 encoding 하기 위해, Multiway Transformers 를 backbone 으로 사용함
  - Figure 2 - 각 Multiway Transfomer block 은 (1) shared self-attention module, (2) 다양한 모달리티를 위한 feed-forwared entworks (i.g. modality experts) 로 구성됨

    (1) Shared self-attention module 은 다른 모달리티 간 alignment 학습을 가능케 하여, multimodal (vision-language) tasks 를 위한 deep fusion 을 가능하게 함

    (2) 각 Input token (images, texts, image-text pairs) 을 각 expert 에 통과시킴
    - Vision Expert (V-FFN), Language Expert (L-FFN), Vision-Language experts (VL-FFN)
    - Modality-specific information capture 가능

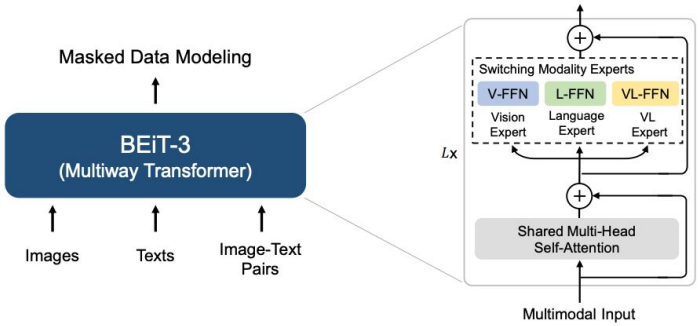


Figure 2: Overview of BEIT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.

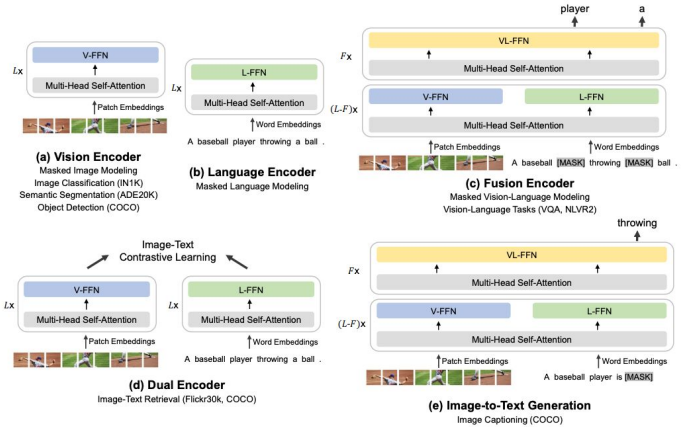


Figure 3: BEIT-3 can be transferred to various vision and vision-language downstream tasks. With a shared Multiway Transformer, we can reuse the model as (a)(b) vision or language encoders; (c) fusion encoders that jointly encode image-text pairs for deep interaction; (d) dual encoders that separately encode modalities for efficient retrieval; (e) sequence-to-sequence learning for image-to-text generation.

BEiT-3: A General-Purpose Multimodal Foundation Model

- **Advance the big convergence from three aspects: 1) <u>Backbone architecture: Multiway Transformers</u>**
  - Figure 3 - 넓은 범위의 downstream task 로 transfer 가능
    - Vision tasks 위한 image backbone : image classification, object detection, instance segmentation, and semantic segmentation
    - I2T (Image-to-Text) retrieval 를 위한 dual encoder 로 fine-tuning
    - Generation Tasks 을 위한 multimodal understanding

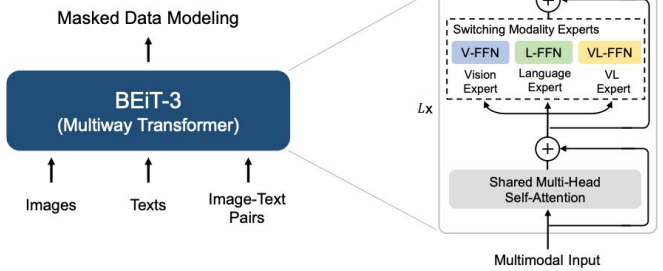


Figure 2: Overview of BEiT-3 pretraining. We perform masked data modeling on monomodal (i.e., images, and texts) and multimodal (i.e., image-text pairs) data with a shared Multiway Transformer as the backbone network.

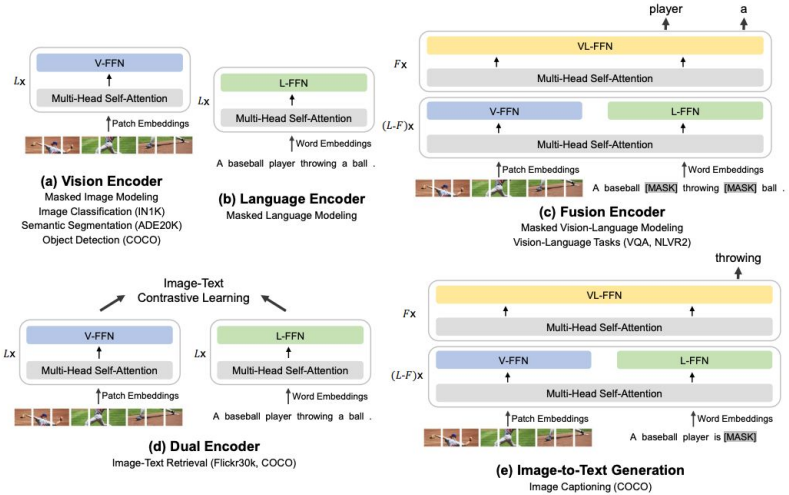


Figure 3: BEiT-3 can be transferred to various vision and vision-language downstream tasks. With a shared Multiway Transformer, we can reuse the model as (a)(b) vision or language encoders; (c) fusion encoders that jointly encode image-text pairs for deep interaction; (d) dual encoders that separately encode modalities for efficient retrieval; (e) sequence-to-sequence learning for image-to-text generation.

- **Advance the big convergence from three aspects: 2) <u>Pretraining Task: Masked Data</u>**
  - Monomodal, Multimodal 데이터를 masked data modeling 을 통해 pretrain 시킴
    - Image patch, text token 의 일부를 mask 한 후, masked token 을 recover 하도록 모델 학습
    - Unified mask-then-predict task 는 (1) representations 을 학습하고, (2) 다른 모달리티 간 alignment 를 학습할 수 있도록 함
  - Implement detail
    - Text data is tokenized by a SentencePiece tokenizer
      - Randomly mask 15% tokens of monomodal texts
      - 50% tokens of texts from image-text paris
    - Image data is tokenized by the tokenizer of BEiT v2
      - Mask 40% of image patches using a block-wise masking strategy as in BEiT

  - 다양한 pretraining tasks (image-text contrast, image-text matching, word-patch/region alignment) 를 수행하는 이전의 vision-language models 과는 다르게, 하나의 pretraining task (mask-then-predict) 만을 수행함
    - 훨씬 작은 pretraining batch size 사용 가능
    - GPU memory cost 함

- **Advance the big convergence from three aspects: 3) <u>Scaling Up: BEiT-3 Pretraining</u>**
  - Backbone Network
    - BEiT-3는 ViT-giant 를 따르는 giant-size foundation model
    - Table 2
      - 40 layer Multiway Transformer with 1408 hidden size, 6144 intermediate size, and 16 attention heads
      - All layers contain both vision experts and language experts. Vision-language experts are employed in the top three Multiway Transformer layers
    - Vision encoder 로 사용할 때는, only vision-related 파라미터만 사용됨!

Model	#Layers	Hidden Size	MLP Size	#Parameters				
				V-FFN	L-FFN	VL-FFN	Shared Attention	Total
BEIT-3	40	1408	6144	692M	692M	52M	317M	1.9B

Table 2: Model configuration of BEIT-3. The architecture layout follows ViT-giant [ZKHB21].

- **Advance the big convergence from three aspects: 3) <u>Scaling Up: BEiT-3 Pretraining</u>**
  - Pretraining Data

Data	Source	Size
Image-Text Pair	CC12M, CC3M, SBU, COCO, VG	21M pairs
Image	ImageNet-21K	14M images
Text	English Wikipedia, BookCorpus, OpenWebText, CC-News, Stories	160GB documents

Table 3: Pretraining data of BEIT-3. All the data are academically accessible.

  - Pretraining Settings
    - Pretrain BEiT-3 for 1M steps
    - Each batch contains 6144 samples
      - 2048 images, 2048 texts and 2048 image-text pairs
    - 14 * 14 patch size and is pretrained at resolution 224 * 224
    - Same image augmentation as in BEiT
      - Random resized cropping, horizontal flipping, and color jittering
    - SentencePiece tokenizer
      - 64k vocab size
    - AdamW optimizer
    - Cosine learning rate decay scheduler with a peak learning rate of 1e-3 and a linear warmup of 10k steps
    - Weight decay is 0.05m Stochastic depth with a rate of 0.1

Experiments

- **Major public benchmarks 에서 vision-language, vision tasks 에 대해 평가함**
- **BEiT-3 는 넓은 범위의 vision, vision-language tasks 에서 SoTA 성능 달성**

Category	Task	Dataset	Metric	Previous SOTA	BEIT-3
Vision	Semantic Segmentation	ADE20K	mIoU	61.4 (FD-SwinV2)	**62.8** (+1.4)
	Object Detection Instance Segmentation	COCO COCO	AP AP	63.3 (DINO) 54.7 (Mask DINO)	**63.7** (+0.4) **54.8** (+0.1)
	Image Classification	ImageNet†	Top-1 acc.	89.0 (FD-CLIP)	**89.6** (+0.6)
Vision-Language	Visual Reasoning	NLVR2	Acc.	87.0 (CoCa)	**92.6** (+5.6)
	Visual QA	VQAv2	VQA acc.	82.3 (CoCa)	**84.0** (+1.7)
	Image Captioning	COCO‡	CIDEr	145.3 (OFA)	**147.6** (+2.3)
	Finetuned Retrieval	COCO Flickr30K	R@1	72.5 (Florence) 92.6 (Florence)	**76.0** (+3.5) **94.2** (+1.6)
	Zero-shot Retrieval	Flickr30K	R@1	86.5 (CoCa)	**88.2** (+1.7)

Table 1: Overview of BEIT-3 results on various vision and vision-language benchmarks. We compare with previous state-of-the-art models, including FD-SwinV2 [WHX+22], DINO [ZLL+22], Mask DINO [ZLL+22], FD-CLIP [WHX+22], CoCa [YWV+22], OFA [WYM+22], Florence [YCC+21]. We report the average of top-1 image-to-text and text-to-image results for retrieval tasks. "†" indicates ImageNet results only using publicly accessible resources. "‡" indicates image captioning results without CIDEr optimization.

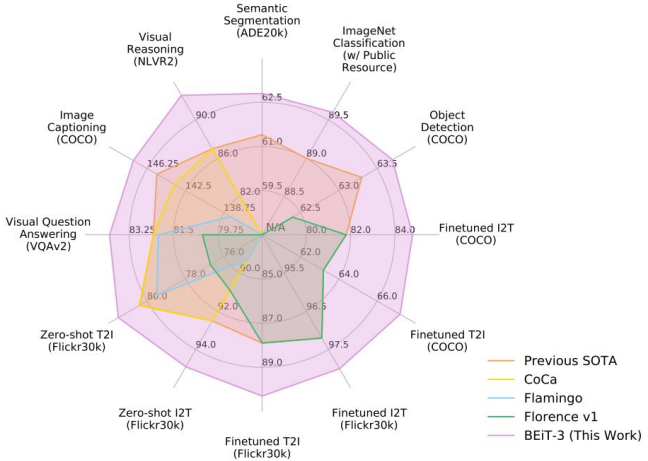


Figure 1: BEIT-3 achieves state-of-the-art performance on a broad range of tasks compared with other customized or foundation models. I2T/T2I is short for image-to-text/text-to-image retrieval.

Experiments

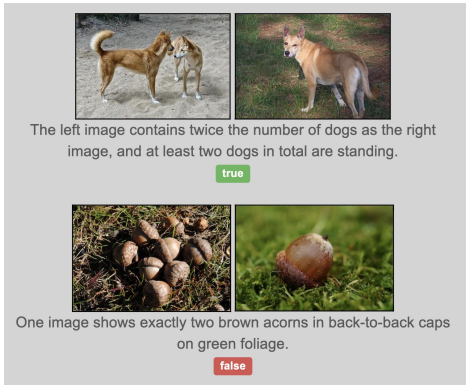
- **Vision-Language Downstream Tasks**
  - **Visual Question Answering (VQA)**
    - 인풋 이미지에 대한 natural language questions 에 대한 답을 하도록 모델 학습 (VQAv2)
    - BEiT-3 outperforms all previous models by a large margin, pushing the SotA to 84.03 with a single model
  - **Visual Reasoning**
    - 인풋 이미지와 텍스트에 대한 reasoning 과 함께 문제를 설명해야 함 (NLVR2)
    - BEiT-3 achieves a new SoTA results for visual reasoning
  - **Image Captioning**
    - 인풋 이미지에 대한 natural language caption 생성 (COCO)
    - BEiT-3 outperforms all previous models trained with cross-entropy loss, creating a new SoTA image captioning result

Visual Question Answering (VQAv2)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Visual Reasoning (NLVR2)



Model	VQAv2		NLVR2		COCO Captioning			
	test-dev	test-std	dev	test-P	B@4	M	C	S
Oscar [LYL+20]	73.61	73.82	79.12	80.37	37.4	30.7	127.8	23.5
VinVL [ZLH+21]	76.52	76.60	82.67	83.98	38.5	30.4	130.8	23.4
ALBEF [LSG+21]	75.84	76.04	82.55	83.14	-	-	-	-
BLIP [LLXH22]	78.25	78.32	82.15	82.24	40.4	-	136.7	-
SimVLM [WYY+21]	80.03	80.34	84.53	85.15	40.6	33.7	143.3	**25.4**
Florence [YCC+21]	80.16	80.36	-	-	-	-	-	-
OFA [WYM+22]	82.00	82.00	-	-	43.9	31.8	145.3	24.8
Flamingo [ADL+22]	82.00	82.10	-	-	-	-	138.1	-
CoCa [YWV+22]	82.30	82.30	86.10	87.00	40.9	**33.9**	143.6	24.7
BEIT-3	**84.19**	**84.03**	**91.51**	**92.58**	**44.1**	32.4	**147.6**	**25.4**

Table 4: Results of visual question answering, visual reasoning, and image captioning tasks. We report *vqa-score* on VQAv2 test-dev and test-standard splits, accuracy for NLVR2 development set and public test set (test-P). For COCO image captioning, we report BLEU@4 (B@4), METEOR (M), CIDEr (C), and SPICE (S) on the Karpathy test split. For simplicity, we report captioning results without using CIDEr optimization.

Experiments

- **Vision-Language Downstream Tasks**
  - **Image-Text Retrieval**
    - 이미지와 텍스트 간 유사도 계산 (COCO, Flicker20K dataset)
    - Image-to-text retrieval, and text-to-image retrieval
    - Dual-encoder DEiT-3 outperforms prior models by a large margin, achieving improvement on COCO top-1 i2t, t2i retrieval, and improvement on Flicker20K top-1 i2t, t2i retrieval
    - BEiT-3 also achieve better performance then previous models on Flicker30K zero-shot retrieval

Model	MSCOCO (5K test set)						Flickr30K (1K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Fusion-encoder models												
UNITER [CLY+20]	65.7	88.6	93.8	52.9	79.9	88.0	87.3	98.0	99.2	75.6	94.1	96.8
VILLA [GCL+20]	-	-	-	-	-	-	87.9	97.5	98.8	76.3	94.2	96.8
Oscar [LYL+20]	73.5	92.2	96.0	57.5	82.8	89.8	-	-	-	-	-	-
VinVL [ZLH+21]	75.4	92.9	96.2	58.8	83.5	90.3	-	-	-	-	-	-
Dual encoder + Fusion encoder reranking												
ALBEF [LSG+21]	77.6	94.3	97.2	60.7	84.3	90.5	95.9	99.8	**100.0**	85.6	97.5	98.9
BLIP [LLXH22]	82.4	95.4	97.9	65.1	86.3	91.8	97.4	99.8	99.9	87.6	97.7	99.0
Dual-encoder models												
ALIGN [JYX+21]	77.0	93.5	96.9	59.9	83.3	89.8	95.3	99.8	**100.0**	84.9	97.4	98.6
FILIP [YHH+21]	78.9	94.4	97.4	61.2	84.3	90.6	96.6	**100.0**	**100.0**	87.1	97.7	99.1
Florence [YCC+21]	81.8	95.2	-	63.2	85.7	-	97.2	99.9	-	87.9	98.1	-
BEIT-3	**84.8**	**96.5**	**98.3**	**67.2**	**87.7**	**92.8**	**98.0**	**100.0**	**100.0**	**90.3**	**98.7**	**99.5**

Table 5: Finetuning results of image-to-text retrieval and text-to-image retrieval on COCO and Flickr30K. Notice that dual-encoder models are more efficient than fusion-encoder-based models for the retrieval tasks.

Model	Flickr30K (1K test set)					
	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
FLAVA [SHG+21]	67.7	94.0	-	65.2	89.4	-
CLIP [RKH+21]	88.0	98.7	99.4	68.7	90.6	95.2
ALIGN [JYX+21]	88.6	98.7	99.7	75.7	93.8	96.8
FILIP [YHH+21]	89.8	99.2	99.8	75.0	93.4	96.3
Florence [YCC+21]	90.9	99.1	-	76.7	93.6	-
Flamingo [ADL+22]	89.3	98.8	99.7	79.5	95.3	**97.9**
CoCa [YWV+22]	92.5	99.5	99.9	80.4	**95.7**	97.7
BEIT-3	**94.9**	**99.9**	**100.0**	**81.5**	95.6	97.8

Table 6: Zero-shot image-to-text retrieval and text-to-image retrieval on Flickr30K.

Experiments

- **Vision Downstream Tasks**
  - **Object Detection and Instance Segmentation**
    - Fine-tuning on COCO 2017 benchmark (First conduct intermediate fine-tuning on the Objects365 dataset)
    - Use BEiT-3 as the backbone and follow ViTDet
    - Achieves the best results on the COCO test-dev set with a smaller image size used for fine-tuning, reaching up to 63.7 box AP and 54.8 mask AP
  - **Semantic Segmentation**
    - ADE20K dataset
    - Use a dense prediction task adapter and employ Mask2Former as the segmentation framework
    - **Directly follow the task transfer settings of ViT-Adapter**
    - BEiT-3 creates a new SoTA with 62.8 mIoU, outperforming FD-SwinV2 giant model with 3B parames by 1.4 points

Model	Extra OD Data	Maximum Image Size	COCO test-dev AP^{box}	COCO test-dev AP^{mask}
ViT-Adapter [CDW+22]	-	1600	60.1	52.1
DyHead [DCX+21]	ImageNet-Pseudo Labels	2000	60.6	-
Soft Teacher [XZH+21]	Object365	-	61.3	53.0
GLIP [LZZ+21]	FourODs	-	61.5	-
GLIPv2 [ZZH+22]	FourODs	-	62.4	-
Florence [YCC+21]	FLOD-9M	2500	62.4	-
SwinV2-G [LHL+21]	Object365	1536	63.1	54.4
Mask DINO [LZX+22]	Object365	1280	-	54.7
DINO [ZLL+22]	Object365	2000	63.3	-
BEiT-3	Object365	1280	**63.7**	**54.8**

Model	Crop Size	ADE20K mIoU	ADE20K +MS
HorNet [RZT+22]	640^2	57.5	57.9
SeMask [JSO+21]	640^2	57.0	58.3
SwinV2-G [LHL+21]	896^2	59.3	59.9
ViT-Adapter [CDW+22]	896^2	59.4	60.5
Mask DINO [LZX+22]	-	59.5	60.8
FD-SwinV2-G [WHX+22]	896^2	-	61.4
BEiT-3	896^2	**62.0**	**62.8**

Experiments

- **Vision Downstream Tasks**
  - **Image Classification**
    - Evaluate the model on ImageNet-1K
    - Vision encoder 에 task layer (classification task layer) 를 추가하지 않고, i2t retrieval task 로 구성함
      - Category name 을 text 로 사용하여 image-text pair 구성
      - 이미지에 가장 관련있는 label 를 찾도록 dual encoder 로 학습 (i2t retrieval)
    - 추론 시, 가능한 class feature embedding 을 계산하고, image feature embedding 을 계산하여, cosine similarity scores 를 계산해 가장 유사한 값을 가진 class로 classification

Model	Extra Data	Image Size	ImageNet
*With extra **private** image-tag data*			
SwinV2-G [LHL+21]	IN-22K-ext-70M	640^2	90.2
ViT-G [ZKHB21]	JFT-3B	518^2	90.5
CoAtNet-7 [DLLT21]	JFT-3B	512^2	90.9
Model Soups [WIG+22]	JFT-3B	500^2	91.0
CoCa [YWV+22]	JFT-3B	576^2	91.0
*With only **public** image-tag data*			
BEiT [BDPW22]	IN-21K	512^2	88.6
CoAtNet-4 [DLLT21]	IN-21K	512^2	88.6
MaxViT [TTZ+22]	IN-21K	512^2	88.7
MViTv2 [LWF+22]	IN-21K	512^2	88.8
FD-CLIP [WHX+22]	IN-21K	336^2	89.0
BEiT-3	IN-21K	336^2	**89.6**

Table 9: Top-1 accuracy on ImageNet-1K.

of possible class names and the feature embedding of the image. Their cosine similarity scores are then calculated to predict the most probable label for each image. Table 9 reports the results on ImageNet-1K. We first perform intermediate finetuning on ImageNet-21K, then we train the model on ImageNet-1K. For a fair comparison, we compare with the previous models only using public image-tag data. BEiT-3 outperforms prior models, creating a new state-of-the-art result when only using public image-tag data.

Vision Transformer Adapter for Dense Predictions

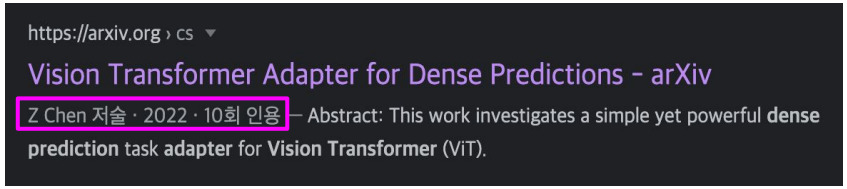
VISION TRANSFORMER ADAPTER FOR DENSE PREDICTIONS

Zhe Chen[1,2], **Yuchen Duan**[3,2], **Wenhai Wang**[2], **Junjun He**[2], **Tong Lu**[1], **Jifeng Dai**[2,3], **Yu Qiao**[2]
[1]Nanjing University, [2]Shanghai AI Laboratory, [3]Tsinghua University

ABSTRACT

This work investigates a simple yet powerful dense prediction task adapter for Vision Transformer (ViT). Unlike recently advanced variants that incorporate vision-specific inductive biases into their architectures, the plain ViT suffers inferior performance on dense predictions due to weak prior assumptions. To address this issue, we propose the ViT-Adapter, which allows plain ViT to achieve comparable performance to vision-specific transformers. Specifically, the backbone in our framework is a plain ViT that can learn powerful representations from large-scale multi-modal data. When transferring to downstream tasks, a **pre-training-free adapter** is used to introduce the image-related inductive biases into the model, making it suitable for these tasks. We verify ViT-Adapter on multiple dense prediction tasks, including object detection, instance segmentation, and semantic segmentation. Notably, without using extra detection data, our ViT-Adapter-L yields state-of-the-art **60.9** box AP and **53.0** mask AP on COCO test-dev. We hope that the ViT-Adapter could serve as an alternative for vision-specific transformers and facilitate future research. The code and models will be released at https://github.com/czczup/ViT-Adapter.

https://arxiv.org/abs/2205.08534
https://github.com/czczup/ViT-Adapter



Vision Transformer Adapter for Dense Predictions

- Dense prediction task adapter for ViT
  - 최근 vision-specific inductive bias 가 내포된 다양한 아키텍처와는 달리, plain ViT 는 약한 prior assumptions 때문에 dense prediction 성능이 떨어짐
  - 해당 문제를 해결하기 위해서, 본 연구에서 simple and powerful dense prediction 을 위한 ViT-Adapter 를 제안하며, 이는 plain ViT 가 vision-specific transformer 에 대응할만한 성능을 가지도록 함
  - 구체적으로, plain ViT backbone 은 large-scale multi-modal 데이터로 강력한 representations 을 학습하며, downstream task 에서 pre-training-free adapter 를 사용함으로써 image-related inductive biases 를 모델이 학습할 수 있도록 함
  - 다양한 dense prediction (object detection, instance segmentation, semantic segmentation) 에서 성능을 확인했고, ViT-Adapter-L 은 COCO test-dev 에서 60.9 box AP, 53.0 mask AP 달성

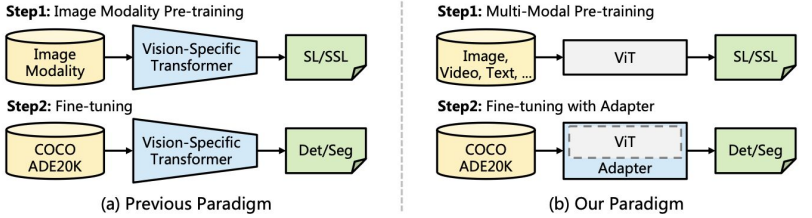


Figure 1: **Previous paradigm *vs.* our paradigm.** (a) Previous paradigm designs vision-specific models and pre-trains on large-scale image datasets via supervised learning (SL) or self-supervised learning (SSL) and then fine-tunes them on downstream tasks. (b) We propose a pre-training-free adapter to close the performance gap between the plain ViT (Dosovitskiy et al., 2020) and vision-specific transformers (*e.g.*, Swin Transformer (Liu et al., 2021b), PVT (Wang et al., 2021)) for dense prediction tasks. Compared to the previous paradigm, our method preserves the flexibility of ViT and thus could benefit from advanced multi-modal pre-training.

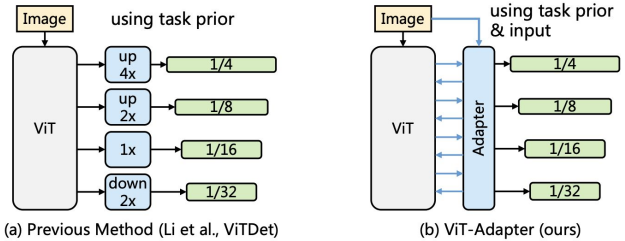
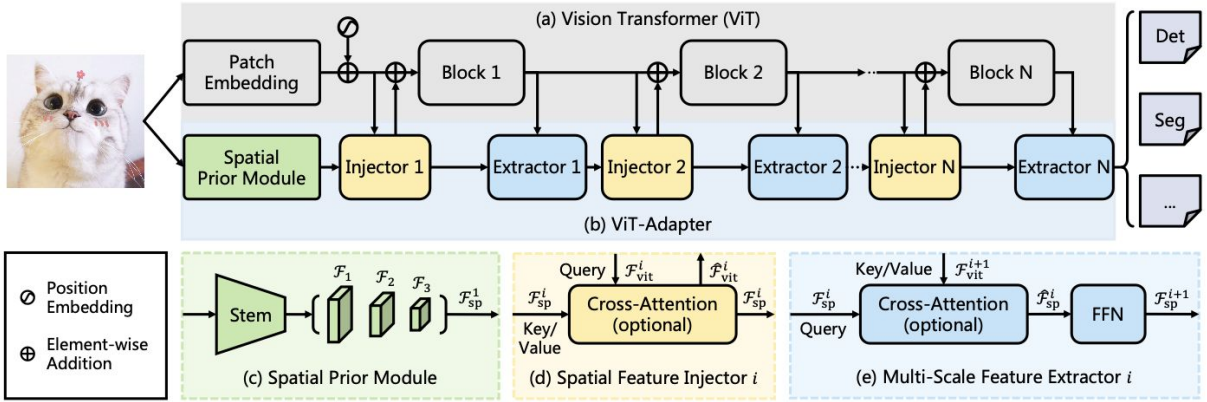


Figure 3: **Overview of our ViT-Adapter and two related approaches.** Li et al. (2021b) and ViTDet (Li et al., 2022b) build simple feature pyramid to adapt plain ViT for object detection, which only consider task prior. Differently, our adapter utilizes both task prior and the input image.

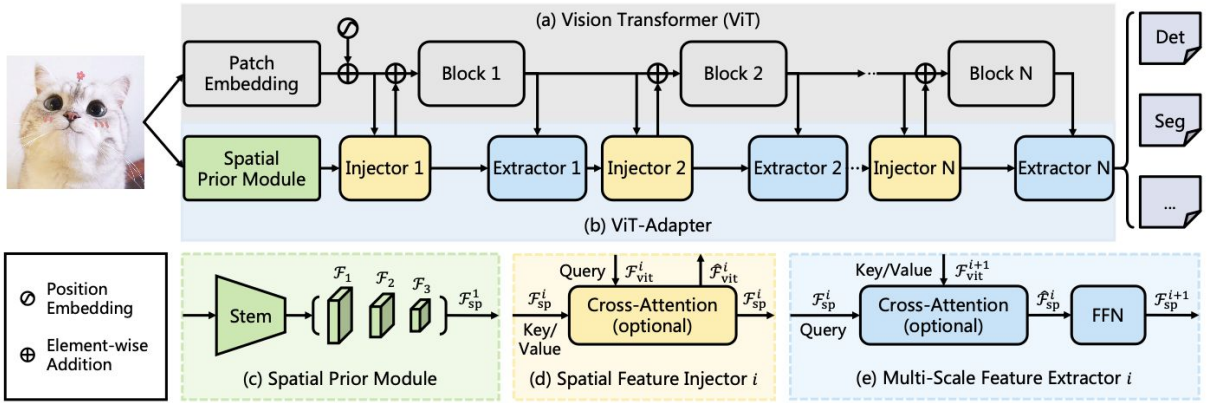
Vision Transformer Adapter for Dense Predictions

- Vision Transformer Adapter
  - 2개 파트로 구성됨
    - Plain ViT
      - L transformer encoder layers
    - ViT-Adapter
      - (1) A spatial prior module to capture spatial features from the input image
      - (2) A spatial feature injector in inject spatial priors into the ViT
      - (3) A multi-scale feature extractor to extract hierarchical features from the single-scale features of ViT



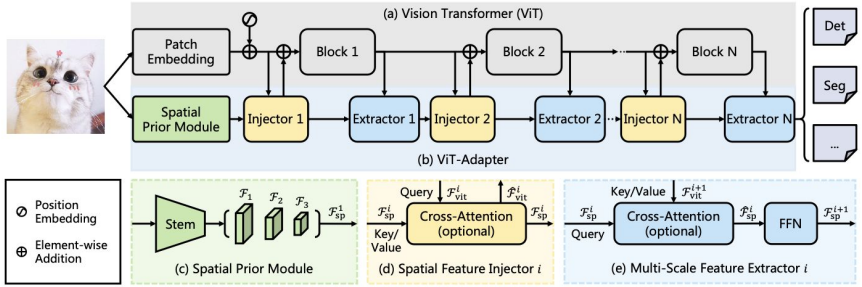
Vision Transformer Adapter for Dense Predictions

- Vision Transformer Adapter
  - 2개 파트로 구성됨
    - Plain ViT
      - Input image 는 16*16 non-overlapping patch 로 쪼개져 patch embedding
      - 각 patch 들은 D-dimensional tokens 으로 flatten, project 되고, feature resolution 은 원본 이미지의 1/16 으로 축소
      - 각 token 들에 positional embedding 추가되고, L 개의 encoder layer 거침



Vision Transformer Adapter for Dense Predictions

- Vision Transformer Adapter
  - 2개 파트로 구성됨
    - ViT-Adapter
      - (1) **A spatial prior module** to capture spatial features from the input image
        - Input 이미지를 Spatial Prior Module 에 입력
        - 3개의 타겟 resolutions (⅛, 1/16, 1/32) 의 spatial features 추출
        - Feature interaction 으로 사용되기 위해서, flatten + concatenated 됨
      - (2) **A spatial feature injector** in inject spatial priors into the ViT
        - Input feature 를 query로, spatial feature 를 key, value 로 사용하여 spatial feature 를 ViT input feature 로 inject 하기 위해서 cross-attention 수행
        - $$\mathcal{F}^i_{\text{vit}} = \mathcal{F}^i_{\text{vit}} + \gamma^i \text{Attention}(\text{norm}(\mathcal{F}^i_{\text{vit}}), \text{norm}(\mathcal{F}^i_{\text{sp}})),$$



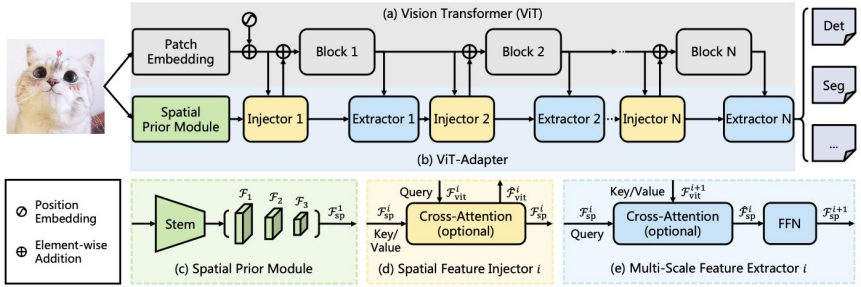
where the $\text{norm}(\cdot)$ is LayerNorm (Ba et al., 2016), and the attention layer $\text{Attention}(\cdot)$ suggests using sparse attention. In addition, we apply a learnable vector $\gamma^i \in \mathbb{R}^D$ to balance the attention layer's output and the input feature $\mathcal{F}^i_{\text{vit}}$, which is initialized with $\mathbf{0}$. This initialization strategy ensures that the feature distribution of $\mathcal{F}^i_{\text{vit}}$ will not be modified drastically due to the injection of spatial priors, thus making better use of the pre-trained weights of ViT.

Vision Transformer Adapter for Dense Predictions

- Vision Transformer Adapter
  - 2개 파트로 구성됨
    - ViT-Adapter
      - **A multi-scale feature extractor** to extract hierarchical features from the single-scale features of ViT
        - Spatial priors 를 ViT 에 injecting 한 후, ViT 를 거친 output feature 를 얻음 (key/value)
        - **multi-scale feature extractor (**Cross-attention layer 와 feed-forward network (FFN)) 을 거쳐서 multi-scale feature 추출

$$\mathcal{F}_{\text{sp}}^{i+1} = \hat{\mathcal{F}}_{\text{sp}}^{i} + \text{FFN}(\text{norm}(\hat{\mathcal{F}}_{\text{sp}}^{i})),$$

$$\hat{\mathcal{F}}_{\text{sp}}^{i} = \mathcal{F}_{\text{sp}}^{i} + \text{Attention}(\text{norm}(\mathcal{F}_{\text{sp}}^{i}), \text{norm}(\mathcal{F}_{\text{vit}}^{i+1})),$$



in which we use the spatial feature $\mathcal{F}_{\text{sp}}^{i} \in \mathbb{R}^{(\frac{HW}{8^2}+\frac{HW}{16^2}+\frac{HW}{32^2})\times D}$ as the query, and the output feature $\mathcal{F}_{\text{vit}}^{i+1} \in \mathbb{R}^{\frac{HW}{16^2}\times D}$ as the key and value for cross-attention. As same as the spatial feature injector, we adopt sparse attention here to reduce computational cost. The generated spatial feature $\mathcal{F}_{\text{sp}}^{i+1}$ will be used as the input of the next spatial feature injector.

Vision Transformer Adapter for Dense Predictions

- Vision Transformer Adapter
  - 아키텍처 configurations
  - 4 different sizes of ViT
    - ViT-T (2.5M), ViT-S (5.8M), ViT-B (14.0M), and ViT-L (23.7M)

Variants	Settings of ViT					Settings of Adapter				Total
	Layers	Width	FFN	Heads	#Param	N	FFN	Heads	#Param	Param
Tiny (T)	12	192	768	3	5.5M	4	48	6	2.5M	8.0M
Small (S)	12	384	1536	6	21.7M	4	96	6	5.8M	27.5M
Base (B)	12	768	3072	12	85.8M	4	192	12	14.0M	99.8M
Large (L)	24	1024	4096	16	303.3M	4	256	16	23.7M	327.0M

Table 10: **Configurations of the ViT-Adapter.** We apply our adapters on four different settings of ViT, including ViT-T, ViT-S, ViT-B, and ViT-L, covering a wide range of different model sizes.

Experiments

- Object detection and Instance segmentation
  - Based on MMDetection and the COCO dataset
  - Use 4 mainstream detectors to evaluate ViT-Adapter, including Mask R-CNN, Cascade Mask R-CNN, ATSS, and GFL

Mask R-CNN

Method	#Param (M)	Mask R-CNN 1× schedule						Mask R-CNN 3×+MS schedule					
		AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}	AP^b	AP^b_{50}	AP^b_{75}	AP^m	AP^m_{50}	AP^m_{75}
PVT-Tiny (Wang et al., 2021)	32.9	36.7	59.2	39.3	35.1	56.7	37.3	39.8	62.2	43.0	37.4	59.3	39.9
PVTv2-B1 (Wang et al., 2022a)	33.7	41.8	64.3	45.9	38.8	61.2	41.6	44.9	67.3	49.4	40.8	64.0	43.8
ViT-T (Li et al., 2021b)	26.1	35.5	58.1	37.8	33.5	54.9	35.1	40.2	62.9	43.5	37.0	59.6	39.0
ViTDet-T (Li et al., 2022b)	26.6	35.7	57.7	38.4	33.5	54.7	35.2	40.4	63.3	43.9	37.1	60.1	39.3
ViT-Adapter-T (ours)	28.1	41.1	62.5	44.3	37.5	59.7	39.9	46.0	67.6	50.4	41.0	64.4	44.1
PVT-Small (Wang et al., 2021)	44.1	40.4	62.9	43.8	37.8	60.1	40.3	43.0	65.3	46.9	39.9	62.5	42.8
PVTv2-B2 (Wang et al., 2022a)	45.0	45.3	67.1	49.6	41.2	64.2	44.4	47.8	69.7	52.6	43.1	66.8	46.7
Swin-T (Liu et al., 2021b)	47.8	42.7	65.2	46.8	39.3	62.2	42.2	46.0	68.1	50.3	41.6	65.1	44.9
ConvNeXt-T (Liu et al., 2022)	48.1	44.2	66.6	48.3	40.1	63.3	42.8	46.2	67.9	50.8	41.7	65.0	44.9
Focal-T (Yang et al., 2021)	48.8	44.8	67.7	49.2	41.0	64.7	44.2	47.2	69.4	51.9	42.7	66.5	45.9
ViT-S (Li et al., 2021b)	43.8	40.2	63.1	43.4	37.1	59.9	39.3	44.0	66.9	47.8	39.9	63.4	42.2
ViTDet-S (Li et al., 2022b)	45.7	40.6	63.3	43.5	37.1	60.0	38.8	44.5	66.9	48.4	40.1	63.6	42.5
ViT-Adapter-S (ours)	47.8	44.7	65.8	48.3	39.9	62.5	42.8	48.2	69.7	52.5	42.8	66.4	45.9
PVTv2-B5 (Wang et al., 2022a)	101.6	47.4	68.6	51.9	42.5	65.7	46.0	48.4	69.2	52.9	42.9	66.6	46.2
Swin-B (Liu et al., 2021b)	107.1	46.9	-	-	42.3	-	-	48.6	70.0	53.4	43.3	67.1	46.7
ViT-B (Li et al., 2021b)	113.6	42.9	65.7	46.8	39.4	62.6	42.0	45.8	68.2	50.1	41.3	65.1	44.4
ViTDet-B (Li et al., 2022b)	121.3	43.2	65.8	46.9	39.2	62.7	41.4	46.3	68.6	50.5	41.6	65.3	44.5
ViT-Adapter-B (ours)	120.2	47.0	68.2	51.4	41.8	65.1	44.9	49.6	70.6	54.0	43.6	67.7	46.9
ViT-L† (Li et al., 2021b)	337.3	45.7	68.9	49.4	41.5	65.6	44.6	48.3	70.4	52.9	43.4	67.9	46.6
ViTDet-L† (Li et al., 2022b)	350.1	46.2	69.2	50.3	41.4	65.8	44.1	49.1	71.5	53.8	44.0	68.5	47.6
ViT-Adapter-L† (ours)	347.9	48.7	70.1	53.2	43.3	67.0	46.9	52.1	73.8	56.5	46.0	70.5	49.7

Table 1: **Object detection and instance segmentation with Mask R-CNN on COCO val2017.** For fair comparison, we initialize all ViT-T/S/B models with the regular ImageNet-1K pre-training (Touvron et al., 2021), and ViT-L† with the ImageNet-22K weights from (Steiner et al., 2021).

Cascade Mask R-CNN, ATSS, and GFL

Method	AP^b	AP^b_{50}	AP^b_{75}	#P
Cascade Mask R-CNN 3×+MS schedule				
Swin-T (Liu et al., 2021b)	50.5	69.3	54.9	86M
Shuffle-T (Huang et al., 2021b)	50.8	69.6	55.1	86M
PVTv2-B2 (Wang et al., 2022a)	51.1	69.8	55.3	83M
Focal-T (Yang et al., 2021)	51.5	70.6	55.9	87M
ViT-S (Li et al., 2021b)	47.9	67.1	51.7	82M
ViT-Adapter-S (ours)	51.5	70.1	55.8	86M
Swin-B (Liu et al., 2021b)	51.9	70.9	57.0	145M
Shuffle-B (Huang et al., 2021b)	52.2	71.3	57.0	145M
ViT-B (Li et al., 2021b)	50.1	69.3	54.3	151M
ViT-Adapter-B (ours)	52.1	70.6	56.5	158M

Method	AP^b	AP^b_{50}	AP^b_{75}	#P
ATSS 3×+MS schedule				
Swin-T (Liu et al., 2021b)	47.2	66.5	51.3	36M
Focal-T (Yang et al., 2021)	49.5	68.8	53.9	37M
PVTv2-B2 (Wang et al., 2022a)	49.9	69.1	54.1	33M
ViT-S (Li et al., 2021b)	45.2	64.8	49.0	32M
ViT-Adapter-S (ours)	49.6	68.5	54.0	36M
GFL 3×+MS schedule				
Swin-T (Liu et al., 2021b)	47.6	66.8	51.7	36M
PVTv2-B2 (Wang et al., 2022a)	50.2	69.4	54.7	33M
ViT-S (Li et al., 2021b)	46.0	65.5	49.7	32M
ViT-Adapter-S (ours)	50.0	69.1	54.3	36M

Table 2: **Object detection with different frameworks on COCO val2017.** For fair comparison, we initialize all ViT-S/B models with the regular ImageNet-1K pre-training (Touvron et al., 2021). "#P" denotes the number of parameters. "MS" means multi-scale training.

Experiments

- Semantic segmentation
  - Evaluate ViT-Adapter on semantic segmentation with the ADE20K dataset and MMSegmentation codebase
  - Semantic FPN and UperNet are employed as the basic frameworks
  - Multi-Modal Pre-training
    - Apply the multi-modal pre-trained weights from Uni-Perceiver (Zhu et al., 2021) for semantic segmentation. As shown in Table 4, for Semantic FPN and UperNet, replacing the ImageNet-22K pre-training with multi-modal pre-training benefits our ViT-Adapter-LF with impressive gains of 1.3 mIoU and 1.6 mIoU, respectively.

Method	Pre-train	Crop Size	Semantic FPN 80k			UperNet 160k		
			#Param	mIoU	+MS	#Param	mIoU	+MS
PVT-Tiny (Wang et al., 2021)	IN-1K	512×512	17.0M	36.6	37.3	43.2M	38.5	39.0
ViT-T (Li et al., 2021b)	IN-1K	512×512	10.2M	39.4	40.5	34.1M	41.7	42.6
ViT-Adapter-T (ours)	IN-1K	512×512	12.2M	41.7	42.1	36.1M	42.6	43.6
PVT-Small (Wang et al., 2021)	IN-1K	512×512	28.2M	41.9	42.3	54.5M	43.7	44.0
PVTv2-B2 (Wang et al., 2022a)	IN-1K	512×512	29.1M	45.2	45.7	-	-	-
Swin-T (Liu et al., 2021b)	IN-1K	512×512	31.9M	41.5	-	59.9M	44.5	45.8
Twins-SVT-S (Chu et al., 2021a)	IN-1K	512×512	28.3M	43.2	-	54.4M	46.2	47.1
ViT-S (Li et al., 2021b)	IN-1K	512×512	27.8M	44.6	45.8	53.6M	44.6	45.7
ViT-Adapter-S (ours)	IN-1K	512×512	31.9M	46.1	46.6	57.6M	46.2	47.1
Swin-B (Liu et al., 2021b)	IN-1K	512×512	91.2M	46.0	-	121.0M	48.1	49.7
Twins-SVT-L (Chu et al., 2021a)	IN-1K	512×512	103.7M	46.7	-	133.0M	48.8	50.2
ViT-B (Li et al., 2021b)	IN-1K	512×512	98.0M	46.4	47.6	127.3M	46.1	47.1
ViT-Adapter-B (ours)	IN-1K	512×512	104.6M	47.9	48.9	133.9M	48.8	49.7
Swin-B† (Liu et al., 2021b)	IN-22K	640×640	-	-	-	121.0M	50.0	51.7
Swin-L† (Liu et al., 2021b)	IN-22K	640×640	-	-	-	234.0M	52.1	53.5
ViT-Adapter-B† (ours)	IN-22K	512×512	104.6M	50.7	51.9	133.9M	51.9	52.5
ViT-Adapter-L† (ours)	IN-22K	512×512	332.0M	52.9	53.7	363.8M	53.4	54.4
ViT-Adapter-L* (ours)	MM	512×512	332.0M	54.2	54.7	363.8M	55.0	55.4

Table 4: **Semantic segmentation on the ADE20K val set.** Semantic FPN (Kirillov et al., 2019) and UperNet (Xiao et al., 2018) are used as segmentation frameworks. "IN-1K/22K" and "MM" represent ImageNet-1K/22K and multi-modal pre-training. "MS" denotes multi-scale testing.

Experiments

- Comparison with SoTA
  - SoTA detection, segmentation frameworks 와 결합하여 ViT-Adapter 성능 평가
  - Plain ViT detector 가 유의미한 성능 향상을 하였음

Results. As shown in Table 5, our method reaches state-of-the-art performance. While these results may be partly due to the effectiveness of advanced pre-training techniques, our study demonstrates that plain ViT detectors and segmenters can challenge the entrenched position of hierarchical backbones.

COCO test-dev	AP^b	AP^m
CB-Swin-L	60.1	52.3
SwinV2-L	60.8	52.7
ViT-Adapter-L	**60.9**	**53.0**

ADE20K val	mIoU
FD-SwinV2-G	61.4
ViT-Adapter-L	**61.5**
BEiT3(w/ ViT-Adapter)	**62.8**

Table 5: **Comparison with previous SOTA.**

Conclusions

- **BEiT-3: A General-Purpose Multimodal Foundation Model**
  - General-purpose multimodal foundation model 제안
  - 넓은 범위의 vision, vision-language benchmarks 에서 SoTA 성능
  - Key idea
    - Backbone 아키텍처
    - Pretraining task
    - Scaling up
  - Simple and effective
  - For future work, including more modalities (e.g. audio)

- **Vision Transformer Adapter for Dense Predictions**
  - Dense prediction task 에서의 plain ViT 와 vision-specific transformer 성능 gap 을 줄이기 위한 새로운 파라다임 제시
  - Inherent 아키텍처 변경 없이, image-related inductive biases 를 ViT 에 inject 하고 reconstruct fine-grained multi-scale features
  - Object detection, instance segmentation, semantic segmentation 에서 vision-specific transformers 보다 좋은 성능

End of the Document