EVA: Exploring the Limits of Masked Visual Representation Learning at Scale

Yuxin Fang[1,2†] Wen Wang[3,2†] Binhui Xie[4,2†] Quan Sun[2] Ledell Wu[2]
Xinggang Wang[1‡] Tiejun Huang[2] Xinlong Wang[2‡] Yue Cao[2‡]

[1]Huazhong University of Science and Technology [2]Beijing Academy of Artificial Intelligence
[3]Zhejiang University [4]Beijing Institute of Technology

Code & Models: baaivision/EVA/01

Abstract

We launch EVA, a vision-centric foundation model to Explore the limits of Visual representation at scAle using only publicly accessible data. EVA is a vanilla ViT pre-trained to reconstruct the masked out image-text aligned vision features conditioned on visible image patches. Via this pretext task, we can efficiently scale up EVA to one billion parameters, and sets new records on a broad range of representative vision downstream tasks, such as image recognition, video action recognition, object detection, instance segmentation and semantic segmentation without heavy supervised training. Moreover, we observe quantitative changes in scaling EVA result in qualitative changes in transfer learning performance that are not present in other models. For instance, EVA takes a great leap in the challenging large vocabulary instance segmentation task: our model achieves almost the same state-of-the-art performance on LVIS dataset with over a thousand categories and COCO dataset with only eighty categories. Beyond a pure vision encoder, EVA can also serve as a vision-centric, multi-modal pivot to connect images and text. We find initializing the vision tower of a giant CLIP from EVA can greatly stabilize the training and outperform the training from scratch counterpart with much fewer samples and less compute, providing a new direction for scaling up and accelerating the costly training of multi-modal foundation models.

1. Introduction

Scaling up pre-trained language models (PLMs) [9,64,76] has revolutionized natural language processing (NLP) in the past few years. The key to this success lies in the simple and scalable self-supervised learning task of masked signal prediction [29, 74], with which Transformer models [99] could be scaled up to billions of parameters using nearly unlimited unlabelled data, and generalize well to a wide range of downstream tasks with little tuning. With further scaling on compute, data, and model sizes, PLMs have led to not only continuous performance improvements [50, 75, 76], but also a surprising emergence of in-context learning capability [9, 25, 104, 105].

Motivated by the success of model scaling in NLP, it is appealing that we can also translate this success from language to vision, *i.e.*, to scale up a vision-centric foundation model that is beneficial for both vision & multi-modal downstream tasks. Recently, masked image modeling (MIM) [5, 39, 113] has boomed as a viable approach for vision model pre-training and scaling. However, the most competitive billion-sized vision pre-trained models [31, 65, 71, 119] still heavily rely on supervised or weakly-supervised training with hundreds of millions of (often publicly inaccessible) labeled data. MIM is somewhat only adopted as an initialization stage before the heavily supervised pre-training [65], or a pure MIM pre-trained model could not achieve favorable performance at billion-scale model sizes [114]. We regard this gap stems from the fact that natural images are raw and information-sparse. Meanwhile, an ideal vision pretext task needs the abstraction of not only the low-level geometry & structure information, but also high-level semantics, which is hardly captured by pixel-level recovery tasks [112].

In this work, we seek a suitable MIM pretext task for large scale vision representation learning and explore its limits at the scale of one billion parameters with tens of millions of unlabeled data. Recently, there are a few trials leveraging the semantic information from image-image or image-text contrastive learning [13, 22, 73] for MIM pre-training [43, 106, 124], which perform fairly well in vision downstream tasks. However, there remains a debate that **(i)** tokenized semantic features could provide better supervision signal for masked modeling in vision [5, 70, 101], and **(ii)** good performances could be also achieved via a simple post-

†Interns at Beijing Academy of Artificial Intelligence (BAAI).
‡Corresponding authors: Yue Cao (caoyue10@gmail.com), Xinlong Wang (xinlong.wang96@gmail.com) and Xinggang Wang (xgwang@hust.edu.cn).

model	image & video classification ()							object detection (det) & instance segmentation (seg)			semantic segmentation	
	IN-1K ft	IN-1K lin	IN-1K zs	avg. zs	K400	K600	K700	COCO det (`test/val`)	COCO seg (`test/val`)	LVIS seg	COCO-Stuff	ADE20K
Florence	-	-	-	-	86.5	87.8	-	62.4 / 62.0	-	-	-	-
SwinV2-G	-	-	-	-	86.8	-	-	63.1 / 62.5	54.4 / 53.7	-	-	59.9
prev. best	89.6[a]	82.3[b]	78.0[c]	73.1[c]	87.8[d]	88.3[e]	80.4[e]	64.5[f] / 64.2[g]	55.4[h] / 54.5[i]	49.2[j]	52.3[k]	62.8[a]
EVA	**89.7**(+0.1)	**86.5**(+4.2)	**78.5**(+0.5)	**75.7**(+2.6)	**89.7**(+1.9)	**89.8**(+1.5)	**82.9**(+2.5)	**64.7 / 64.5**(+0.2/+0.3)	**55.5 / 55.0**(+0.1/+0.5)	**55.0**(+5.8)	**53.4**(+1.1)	**62.3**(-0.5)

Table 1. **Summary of EVA performance on various mainstream vision benchmarks.** EVA is performant compared with previous best / leading approaches. “”: methods / results that only exploit publicly accessible data / academic resources. “ft”: end-to-end fine-tuning. “lin”: linear probing. “zs”: zero-shot classification. “avg. zs”: averaged zero-shot classification performance on 8 image and 4 video datasets with contrastive language-image pre-training. (timestamp: Nov 10, 2022)

methods / results reference. a: BEiT-3 [101], b: iBOT [124], c: Open CLIP-H [47], d: Text4Vis [109], e: MaskFeat [103], f: Group DETRv2 [19], g: FocalNet [116], h: FD-SwinV2-G [107], i: Mask DINO [57], j: LVIS 2021 competition 1st [35], k: ViT-Adapter [23].

distillation process without masked prediction tasks [107]. Through a pilot empirical study, we find that simply using image-text aligned (*i.e.*, CLIP [73]) vision features as the prediction targets in MIM scales up well and achieves satisfactory performances on a broad range of downstream benchmarks. This pre-training task draws the benefits from both the high-level semantic abstraction of image-text contrastive learning as well as the good capture of geometry & structure in masked image modeling, which typically covers the information needed for most visual perception tasks.

Via this MIM pretext task, we can efficiently scale up a vanilla ViT encoder [31], dubbed **EVA**, to one billion parameters with strong visual representations that transfers well to a wide range of downstream tasks. Using 29.6 million public accessible unlabeled images for pre-training, **EVA** sets new records on several representative vision benchmarks, such as image classification on ImageNet-1K [28] (89.7% top-1 accuracy), object detection and instance segmentation on LVIS [38] (62.2 AP^{box} & 55.0 AP^{mask} on `val`) and COCO [62] (64.5 AP^{box} & 55.0 AP^{mask} on `val`, 64.7 AP^{box} & 55.5 AP^{mask} on `test-dev`), semantic segmentation on COCO-stuff [11] (53.4 $mIoU^{ss}$) and ADE20K [123] (62.3 $mIoU^{ms}$), and video action recognition on Kinetics-400 [51] (89.7% top-1 accuracy), Kinetics-600 [14] (89.8% top-1 accuracy), Kinetics-700 [15] (82.9% top-1 accuracy). Notably, different from other state-of-the-art billion-scale vision foundation models that demand tens of millions of or even billions of labeled images, such as SwinV2-G using ImageNet-21K-ext-70M [65] and ViT-g/G using JFT-3B [119], **EVA** does not need a costly supervised training stage and only leverage images from open-sourced datasets for academic reproducibility.

tokenize? [70]	pt epochs	ImageNet-1K top-1 acc.	ADE20K $mIoU^{ss}$
✗	-	85.0	52.6
✓	300	85.0	52.7
✓	1600	85.5	53.1
✗	800	**85.5**	**53.3**

(a) **(Additional) semantic feature tokenization** is not required for achieving good downstream performance.

distill.? [107]	pt epochs	ImageNet-1K top-1 acc.	ADE20K $mIoU^{ss}$
✗	-	85.0	52.6
✓	300	85.1	52.5
✓	800	85.1	52.7
✗	800	**85.5**	**53.3**

(b) **Feature distillation** fails to achieve consistent performance gain as the pre-training becomes longer.

Table 2. **Pilot experiment**. We evaluate different pre-training approaches using ViT-B and report their performance on ImageNet-1K image classification (top-1 accuracy) and ADE20K semantic segmentation (single-scale mIoU). Numbers in grey refer to the results of directly fine-tuning CLIP vision encoder on corresponding downstream tasks. Default settings for **EVA** pre-training are marked in purple, *i.e.*, directly regressing the masked out CLIP vision features conditioned on visible image patches.

Moreover, we observe quantitative changes in scaling **EVA** result in qualitative changes in transfer learning performance that are not observed in other smaller-scale models, *e.g.*, **EVA** makes a significant breakthrough in the challenging large vocabulary object-level recognition task: our model achieves almost the same performance on LVIS [38], an instance segmentation benchmark with more than 1,200 categories, as COCO [62], which almost shares the same image set as LVIS but with only 80 categories annotated. This emergent ability well matches the expectation of model scaling [105], that larger capability of model results in not only predictable performance improvements on standard benchmarks, but also unpredictable phenomenons and capabilities for resolving more challenging tasks.

Going beyond a pure vision encoder, **EVA** can also serve as a vision-centric, multi-modal pivot that builds a bridge between vision and language. We show that initializing the image encoder via pre-trained **EVA** in a 1.1 billion parameters CLIP model can outperform the training from scratch counterpart on a broad range of zero-shot image / video classification benchmarks with much fewer samples and less compute. Moreover, **EVA** can greatly stabilize the giant CLIP's training & optimization process. Since large CLIP models usually suffer from training instability and inefficiency issues [2, 47], we hope our solution opens up a new direction for scaling up and accelerating the costly training of multi-modal foundation models.

By scaling up vision-centric foundation models with MIM pre-training to achieve strong performance on broad downstream tasks, we hope **EVA** would bridge the gap between vision and language with masked signal modeling, and contributes to the big convergence across different modalities.

patch size	#layers	hidden dim	mlp dim	attn heads	#param.
14×14	40	1408	6144	16	1011M

(a) EVA architecture configurations.

dataset	total size
ImageNet-21K, CC12M, CC3M, Object365, COCO, ADE	29.6M images

(b) datasets for pre-training EVA.

image size	batch size	optimizer	peak lr	(β_1, β_2)	pt epochs
224^2	4096	AdamW	1e-3	(0.9, 0.98)	150

(c) some pre-training settings and hyper-parameters.

precision	ZeRO	#gpus	samples / sec.	max mem.	pt days
`fp16`	stage-1	128	~3150	~26.5GB	~14.5

(d) basic statistics of EVA pre-training.

Table 3. **A brief summary of pre-training settings and configurations for EVA.**

2. Fly EVA to the Moon

We first conduct a series of pilot experiments for choosing an ideal vision pretext task in §2.1, then we scale up EVA pre-training via the chosen pre-training objective in §2.2. Finally, we evaluate the pre-trained representation on various downstream tasks in §2.3. Detailed experimental settings and configurations are in Appendix.

2.1. The Feature Instrumentality Project

In this section, we seek a MIM vision pretext task with compelling transfer performance. Based on previous literature on vision pre-training, we study two promising candidates: **(i)** recovering the masked out *tokenized* semantic vision features [5, 70, 101], and **(ii)** feature *distillation* from strong pre-trained representation as in [107]. Both of them exploit pre-trained image-text aligned vision features (*i.e*., CLIP [73] vision features). Via a series of pilot experiments shown in Table 2, we find that: **(i)** the (additional) CLIP feature tokenization process is unnecessary for achieving good downstream performance **(ii)** feature *distillation* fails to provide consistent performance gain as the pre-training becomes longer. Instead, we find that simply reconstructing the masked out CLIP vision features conditioned on visible image patches is highly performant, which is chosen for scaling up EVA.

We clarify that this MIM pretext task is *not* originally proposed by us. Regressing the masked out image-text aligned vision features for MIM pre-training has been studied in MVP [106] and recently has been revisited by MILAN [43]. In this work, we show that this pretext task can scale up to billion-scale parameters and tens of millions of unlabeled images for vision-centric representation learning *without* **(i)** semantic feature quantization / tokenization [5, 70], and **(ii)** explicitly using image-text paired pre-training data and large corpora as in BEiT-3 [101].

2.2. Pre-training

Architecture. The architecture configurations of EVA are in Table 3a. EVA is a vanilla ViT [31] with 1.0B parameters. The shape of her follows ViT giant [119] and the vision encoder of BEiT-3 [101]. We do not use relative positional embeddings [89] and layer-scale [97] during pre-training.

Pre-training objective. EVA is pre-trained to reconstruct the masked out image-text aligned vision features conditioned on visible image patches. We corrupt the input patches with `[MASK]` tokens, and we use block-wise masking with a masking ratio of 40% following [5, 70, 101]. The target for MIM pre-training is from the publicly available OpenAI CLIP-L/14 vision tower trained on 224×224 pixel images [73]. The output feature of EVA is first normalized [3] and then projected to the same dimension as the CLIP feature via a linear layer. We use negative cosine similarity as the loss function.

Pre-training data. The data we used for pre-training EVA are summarized in Table 3b. For CC12M [16] and CC3M [88] datasets, we only use the image data without captions. For COCO [62] and ADE20K [123] datasets, we only use the train set data. ImageNet-21K [28] and Object365 [87] image data are also used. All these data are publicly accessible. The merged dataset for pre-training has 29.6 million images in total.

Pre-training settings & hyper-parameters. As shown in Table 3c, EVA is optimized via Adam [52] with decoupled weight decay [67] of 0.05. The peak learning rate is 1e-3 and decays according to a cosine learning rate schedule. We employed stochastic depth [45] with a rate of 0.1 for regularization and `RandResizeCrop` (0.2, 1) for data augmentation. Color jitter is not used.

Pre-training infrastructure and statistics. Some basic pre-training statistics are available in Table 3d. The GPU we use is NVIDIA A100-SXM4-40GB. Pre-training code is based on BEiT [5] written in PyTorch [69]. We also adopt `DeepSpeed` optimization library [80] with ZeRO stage-1 optimizer [77] to save memory. We find using `fp16` format with dynamic loss scaling is stable enough during the whole course of pre-training while using `bfloat16` format is unnecessary. Since we use `fp16` precision, EVA can also be pre-trained using 16× NVIDIA 24GB (32GB) GPUs with (without) gradient checkpointing [20].

2.3. Evaluation on Downstream Tasks

In this section, we extensively evaluate pre-trained EVA on several representative benchmarks, such as image classification (§2.3.1), video action recognition (§2.3.2), object detection & instance segmentation (§2.3.3), semantic segmentation (§2.3.4), and contrastive image-text pre-training with zero-shot evaluation (§2.3.5). EVA achieves state-of-the-art performance on a broad range of downstream tasks.

model	#param.	extra labeled data	image size	top-1 acc.
		using *private* labeled data		
SwinV2-G [65]	3.0B	IN-21K-ext-70M	640^2	90.2
ViT-G [119]	1.8B	JFT-3B	518^2	90.5
ViT-g (CoCa) [117]	1.0B	JFT-3B+ALIGN	576^2	91.0
		using *public* labeled data		
CoAtNet-4 [27]	275M	IN-21K (14M)	512^2	88.6
MaxViT-XL [98]	475M	IN-21K (14M)	512^2	88.7
MViTv2-H [61]	667M	IN-21K (14M)	512^2	88.8
FD-CLIP-L [107]	304M	IN-21K (14M)	336^2	89.0
BEiT-3 [101]	2.0B	35M img-txt pairs	336^2	89.6
EVA	1.0B	IN-21K (14M)	336^2	**89.6**
EVA	1.0B	IN-21K (14M)	560^2	**89.7**

Table 4. **Comparisons of image classification performance on ImageNet-1K validation set.** With only publicly available data, EVA creates a new state-of-the-art ImageNet-1K image classification result with a canonical linear classifier.

2.3.1 Image Classification

Datasets. For image classification task, we evaluate EVA on ImageNet-1K (IN-1K) [28] validation set. We also evaluate the robustness & generalization capability of EVA along with our training settings & hyper-parameters using ImageNet-V2 matched frequency (IN-V2) [81], ImageNet-ReaL (IN-ReaL) [7], ImageNet-Adversarial (IN-Adv.) [42], ImageNet-Rendition (IN-Ren.) [41], ImageNet-Sketch (IN-Ske.) [100].

Training Settings. Following the conventional setting [5, 70, 101], we first perform intermediate fine-tuning on ImageNet-21K [28] for 60 epochs with an image resolution of 224^2, then EVA is further fine-tuned on ImageNet-1K training set for 10 epochs. Different from [117, 119] that use multi-head attention pooling and BEiT-3 that exploits an additional pre-trained giant language tower as the image classification task layer, we simply adopt a linear layer as the classifier [31]. Notice that the supervised intermediate fine-tuning consumes only ∼1/5 of the time & compute of the MIM pre-training stage. While for other billion-scale vision models such as SwinV2-G-3B, the supervised training phase costs ∼1.5× resources than the MIM pre-training.

Results. Table 4 compares EVA with some state-of-the-art models on ImageNet-1K validation set. EVA achieves 89.6% top-1 accuracy with 336^2 inputs, comparable to BEiT-3. Using a larger image resolution of 560^2 can further boost the top-1 accuracy to 89.7%. Notice that BEiT-3 treats image classification as an image-to-text retrieval task. Therefore they leverage an additional one billion parameters pre-trained language encoder along with 35 million image-text data (21M pairs from CC12M, CC3M, SBU, COCO, VG and 14M pairs from ImageNet-21K) as well as 160GB text data in total. Meanwhile, we simply use a linear classifier on top of EVA with only ImageNet-21K image-tag data used for additional fine-tuning. With only publicly available data, EVA creates a new state-of-the-art image classification result on ImageNet-1K with a much neater architecture.

Robustness & generalization ability evaluation. We evaluate the robustness and generalization capability of EVA trained with an image size of 336^2 on 6 different ImageNet-1K validation set variants. In Table 5, we compare EVA with some top open-sourced models collected by the `timm` library [108]. Following the evaluation procedure in [39], all these models are first fine-tuned on the original ImageNet-1K training set and then evaluated on different validation sets using the *same* fine-tuned model without further hyper-parameter selection and specialized fine-tuning.

model	IN-1K	IN-V2	IN-ReaL	IN-Adv.	IN-Ren.	IN-Ske.	avg.	Δ↓
ConvNeXt	87.5	77.7	90.5	70.8	67.0	53.7	74.5	13.0
SwinV2	87.5	77.3	90.2	73.9	67.7	52.3	74.8	12.7
MAE	87.8	79.2	90.3	76.7	66.5	50.9	75.2	12.6
DeiT3	87.7	79.1	90.2	79.2	70.6	54.9	77.0	10.7
Eff-L2-NS	88.4	80.5	90.6	84.8	74.7	47.6	77.8	10.6
BEiTv2	88.4	80.1	90.3	76.2	76.4	58.3	78.3	10.1
BEiT	88.6	79.9	90.7	81.7	73.2	56.8	78.5	10.1
EVA	89.6	81.6	90.8	86.2	88.3	67.7	84.0	**5.6**

Table 5. **Robustness & generalization capability evaluation on ImageNet-1K variants.** We test each model on different ImageNet-1K validation sets, without any specialized fine-tuning. "avg.": the averaged top-1 accuracy on 6 different ImageNet-1K validation set variants. "Δ↓": The gap between the averaged top-1 accuracy on 6 variants (*i.e.*, IN-{1K, V2, ReaL, Adv., Ren., Ske.}) and the original ImageNet-1K validation set top-1 accuracy (the lower the better).

As shown in Table 5, EVA is the most competitive one in terms of absolute top-1 accuracies. However, these model various in pre-train data (from ImageNet-1K, ImageNet-21K to JFT-300M), input resolutions (from 224^2 to 800^2), model sizes (from hundreds of millions to one billion parameters) as well as architectures (ConvNets, vanilla & hierarchical ViTs), *etc*. Therefore their absolute accuracies are *not* directly comparable. Instead, we are more interested in the *gap* between the averaged top-1 accuracy on 6 validation set variants and the original ImageNet-1K validation set top-1 accuracy (the lower the better), *i.e.*, we care about whether a model along with its training settings biases towards the original validation set and generalize well on other variants. From this perspective, EVA not only achieves the highest averaged accuracy, but also has the smallest performance gap, which reflects the excellent robustness and generalization ability of EVA.

2.3.2 Video Action Recognition

Datasets. For video action recognition, we evaluate EVA on Kinetics-400 (K-400) [51], Kinetics-600 (K-600) [14] and Kinetics-700 (K-700) [15] benchmarks. We first conduct intermediate fine-tuning on a merged dataset coined Kinetics-722 (K-722) that integrates videos from K-400, K-600 and K-700. We remove leaked as well as repeated videos in both training and validation sets. After this data de-duplicating

Source: `link` (timestamp: Nov 10, 2022). The detailed model configurations are (arch-model_size-img_resolution-data): ConvNeXt-XL-384px-21K [66], SwinV2-L-384px-21K [65], MAE-H-448px-1K [39], DeiT3-L-384px-21K [96], EfficientNet-L2&NS-800px-JFT300M [110], BEiTv2-L-224px-21K [70], BEiT-L-512px-21K [5], EVA-g-336px-merged30M&21K.

model	top-1 accuracy Kinetics-400	Kinetics-600	Kinetics-700
MAE [34]	86.8	-	-
SwinV2-G [65]	86.8	-	-
Florence [118]	86.8	88.0	-
MaskFeat [103]	87.0	88.3	80.4
VideoMAE [95]	87.4	-	-
X-CLIP [68]	87.7	88.3	-
CoVeR [120]	87.2	87.9	78.5
CoCa [117] (frozen)	88.0	88.5	81.1
CoCa [117] (finetuned)	88.9	89.4	82.7
EVA	**89.7**	**89.8**	**82.9**

Table 6. **Video action recognition.** With only publicly available K-400, K-600 and K-700 as video pre-training data, EVA is also quite performant in video action recognition tasks.

process, K-722 has 0.63M training videos in total with 722 action classes. A similar approach is also used in [58].

Training & evaluation settings. EVA processes video data simply via spatial-temporal attention as [34, 95] with no specific architectural adaptation for video related tasks. We first train EVA using K-722 training set for 40 epochs with 8 frames and 224^2 resolution, then we fine-tune EVA on each dataset for only 1 or 2 epochs. We set frame$\times$crop$\times$clip to $16\times3\times4$ for fine-tuning and evaluation for all datasets. The frame resolution is 224^2.

Results. As shown in Table 6, EVA achieves better performance compared with some recent video-specific or large foundation models in video recognition. For reference, directly adapting image-only pre-trained EVA to K-400 without K-722 intermediate fine-tuning can also achieve a very competitive top-1 accuracy of 88.4%.

2.3.3 Object Detection & Instance Segmentation

Datasets. We evaluate the object detection and instance segmentation performance of EVA on both COCO [62] and LVIS [38]. COCO is a widely used object-level recognition benchmark with 80 common object categories. LVIS is an emerging large-vocabulary object-level recognition benchmark, which has more than 1,200 object categories as well as more than 2 million high quality instance segmentation masks (nearly $2\times$ of COCO). Notably, COCO and LVIS almost use the same set of images, and both `train` and `val` split of LVIS have a huge overlap with COCO `train` and `val` split. Meanwhile, COCO has much fewer object categories than LVIS (*i.e.*, 80 *v.s.* 1,200+). Therefore it is meaningful to evaluate models' performance on both COCO and LVIS.

Training & evaluation settings. EVA uses Cascade Mask R-CNN [12] as the detector and adopts the training settings (*e.g.*, LSJ data augmentation [36]) & architecture configurations (*e.g.*, interleaved window & global attention) of ViTDet [60]. Following the common practice [65, 101, 121], we first conduct intermediate fine-tuning for the whole detector using Objects365 [87] dataset with a resolution of 1024^2, then we fine-tune the detector on COCO and LVIS `train` split respectively with 1280^2 inputs.

We report single-scale evaluation and multi-scale evaluation / test-time augmentation (tta) results of EVA for comparison. For COCO, Soft-NMS [8] is also applied. For instance segmentation task, the classification score is calibrated [46] via maskness [102].

The model architecture as well as the hyper-parameters for COCO and LVIS are almost the *same* (*i.e.*, the hyper-parameters are nearly "zero-shot" transferred from COCO to LVIS), except we use federated loss [125] and repeat factor sampling [38] following ViTDet on LVIS.

Results. Perhaps COCO is the most fierce vision benchmark. Table 7 compares EVA with some leading approaches on COCO. Our model creates new state-of-the-art results on both object detection and instance segmentation tasks.

Compared with ViTDet-H [60] that uses Cascade Mask R-CNN [12] as well, EVA shows that with a larger model and better encoder & detector pre-training, the performance can be greatly improved with the same detector.

Compared with FocalNet [116] and Group DETRv2 [19] that choose better-established and highly-optimized DINO detector [121], EVA demonstrates that with sufficient model size, data and pre-training, better performance can be also achieved via the classic R-CNN framework [37]. On the other hand, FocalNet and Group DETRv2 are incapable of instance segmentation due to using DINO.

Compared with SwinV2-Giant [65] and FD-SwinV2-Giant [107] that also adopt a (stronger HTC++ [17]) detector from the R-CNN family but with ~3$\times$ model size of EVA, our approach streamlines the pre-training processes and pulls off a "Giant-killing" act via better representations.

Compared with BEiT-3, EVA shows that is possible to build a state-of-the-art object-level recognition system without exploiting **(i)** semantic feature quantization / tokenization [5, 70], and **(ii)** image-text paired pre-training data and large corpora during pre-training.

Analyzing the performance gap between LVIS and COCO. Evaluating models on *both* COCO and LVIS benchmarks is essential, as they share nearly the same image set but differ in the number of annotated object categories. COCO has only 80 annotated categories, while LVIS annotates over 1,200 object categories, resulting in a long-tail distribution that more closely resembles challenging real-world scenarios [38]. In general, LVIS is considered a much more difficult benchmark than COCO for object-level recognition, with conventional methods typically experiencing a significant performance drop on LVIS.

In Table 8a, we analyze the performance gap between the LVIS and COCO benchmarks for EVA and other state-of-the-art approaches. For previous leading methods, such as ViTDet, the performance gap for AP^{box} is around 8, and

model / method	detector	#param.	pre-training data encoder	pre-training data detector	tta?	COCO val AP^{box}	COCO val AP^{mask}	COCO test-dev AP^{box}	COCO test-dev AP^{mask}
Soft-Teacher [115]	HTC++ [17]	284M	IN-21K (14M)	COCO(unlabeled)+O365	✓	60.7	52.5	61.3	53.0
GLIP [59]	DyHead [26]	≥ 284M	IN-21K (14M)	4ODs+GoldG+Cap12M	✓	60.8	-	61.5	-
GLIPv2 [122]	DyHead [26]	≥ 637M	FLD-900M	merged data[a]	✓	-	-	62.4	-
ViTDet-H [60]	CMask R-CNN [12]	692M	IN-1K (1M)	-	✓	61.3	53.1	-	-
Florence [118]	DyHead [26]	≥ 637M	FLD-900M	merged data[a]	✓	62.0	-	62.4	-
SwinV2-G [65]	HTC++ [17]	≥ 3000M	IN-21K-ext-70M	O365	✓	62.5	53.7	63.1	54.4
DINO [121]	-	218M	IN-21K (14M)	O365	✓	63.2	-	63.3	-
Mask DINO [57]	-	223M	IN-21K (14M)	O365	✓	-	54.5	-	54.7
BEiT-3 [101]	CMask R-CNN [12]	1074M	merged data[b]	O365	✓	-	-	63.7	54.8
FD-SwinV2-G [107]	HTC++ [17]	≥ 3000M	IN-21K-ext-70M	O365	✓	-	-	64.2	55.4
FocalNet [116]	DINO [121]	746M	IN-21K (14M)	O365	✓	64.2	-	64.4	-
Group DETRv2 [19]	DINO [121]	629M	IN-1K (1M)	O365	✓	-	-	64.5	-
EVA	CMask R-CNN [12]	1074M	merged-30M	O365	✗	**64.2**	**55.0**	**64.4**	**55.5**
EVA	CMask R-CNN [12]	1074M	merged-30M	O365	✓	**64.5**	-	**64.7**	-

Table 7. **Object detection & instance segmentation on results COCO dataset.** EVA establishes new state-of-the-art results in object detection and instance segmentation tasks on both COCO `val` and `test-dev` splits with the canonical R-CNN [37] object detection & segmentation framework. "tta" refers to test-time augmentation. (timestamp: Nov 10, 2022)

"merged data[a]": FourODs + INBoxes + GoldG + CC15M + SBU, "merged data[b]": IN-21K (14M) + Image-Text (35M) + Text (160GB).

model	AP^{box} COCO	AP^{box} LVIS	AP^{box} Δ↓	AP^{mask} COCO	AP^{mask} LVIS	AP^{mask} Δ↓
(a) evaluation using COCO & LVIS official annotations respectively						
Copy-Paste [36]	57.0	41.6	15.4	48.9	38.1	10.8
ViTDet-H [60]	61.3	53.4	7.9	53.1	48.1	5.0
prev. best	63.2[a]	53.4[b]	9.8	54.5[c]	49.2[d]	5.3
EVA (single-scale test)	**64.1**	**62.2**	**1.9**	**55.0**	**55.0**	**0.0**
(b) evaluation using LVIS `val-5K` annotations						
EVA (single-scale test)	69.6	68.3	**1.3**	59.6	59.8	**-0.2**

Table 8. **LVIS & COCO performance gap on `val` set.** "prev. best" refers to the best *individual* model / result in each benchmark (a: DINO [121], b: ViTDet-H [60], c: Mask DINO [57], d: 2021 competition 1[st] [35]) "Δ↓": the performance gap between LVIS and COCO (the lower the better).

for AP^{mask}, it is around 5. However, when using the same detector (Cascade Mask R-CNN) and nearly identical settings as those in ViTDet pre-trained via MAE-Huge (ViTDet-H), EVA not only achieves state-of-the-art results on both LVIS and COCO benchmarks simultaneously but also significantly reduces the performance gap between them, particularly for the instance segmentation task. EVA attains the same performance on LVIS and COCO using single-scale evaluation. In comparison with ViTDet-H, we demonstrate that a slightly larger model with stronger representations can greatly improve performance on the challenging large vocabulary instance segmentation benchmark, with one *caveat* described below.

Note that the Merged-30M unlabeled images include 15K out of 20K LVIS `val` set images (the Merged-30M images contain all the COCO training images, and the LVIS validation split also includes 15k images from the COCO training set). Although a recent study [33] shows that including unlabeled images from the development / test set for MIM pre-training has minimal impact on the final performance, we conduct a more rigorous analysis of the LVIS and COCO performance gap to eliminate potential data contamination issues: We evaluate both COCO and LVIS models using the 5K images present in both the COCO and LVIS `val` sets, denoted as LVIS `val-5K`. The COCO results are measured using the 80-category COCO subset of LVIS with the higher-quality LVIS annotations (a similar approach also employed in [53], but for a different purpose). The results are shown in Table 8b, and we find that the conclusion remains unchanged.

model	crop size	ADE20K $mIoU^{ss}$	ADE20K $mIoU^{ms}$	COCO-Stuff $mIoU^{ss}$
HorNet [79]	640^2	57.5	57.9	-
SeMask [48]	640^2	57.0	58.3	-
SwinV2-G [65]	896^2	59.3	59.9	-
Mask DINO [57]	896^2	59.5	60.8	-
FD-SwinV2-G [107]	896^2	-	61.4	-
ViT-Adapter [23]	896^2	61.2	61.5	52.3
BEiT-3 [101]	896^2	62.0	62.8	-
EVA	896^2	**61.5**	**62.3**	**53.4**

Table 9. **Semantic segmentation performance on ADE20K and COCO-Stuff-164K dataset.** "$mIoU^{ss}$": mIoU of single-scale evaluation, "$mIoU^{ms}$": mIoU using multi-scale evaluation.

2.3.4 Semantic Segmentation

Dataset. We evaluate EVA on ADE20K [123] and COCO-Stuff-164K [11] datasets for semantic segmentation task. ADE20K includes 150 semantic categories, and has 20k images for training & 2k images for validation. COCO-Stuff-164K augments 164K complex images from COCO with pixel-level annotations that span over 172 categories including 80 things, 91 stuff, and 1 unlabeled class. Compared with ADE20K, COCO-Stuff is a more challenging but under-explored semantic segmentation benchmark.

Training & evaluation settings. We follow the task transfer pipelines of ViT-Adapter [23]+mask2former [24] but with a weakened model adaptation processes due to GPU memory

model	precision	total #param.	image #param.	text #param.	clip training data	samples seen	image size	patch size	batch size	gpus for training
OpenAI CLIP-L	`float16`	430M	304M	124M	CLIP-400M [73]	12B	224^2	14×14	32k	256×V100 (32GB)
ALIGN	`bfloat16`	834M	480M	354M	ALIGN-1.8B [73]	22B	289^2	-	16k	1024×TPUv3
Open CLIP-H	`bfloat16`	1.0B	632M	354M	LAION-2B [85]	32B	224^2	14×14	79k	824×A100 (40GB)
Open CLIP-g	`bfloat16`	1.3B	1.0B	354M	LAION-2B [85]	12B	224^2	14×14	64k	800×A100 (40GB)
EVA CLIP-g	`float16`	1.1B	1.0B	124M	LAION-400M [86]	11B	224^2	14×14	41k	256×A100 (40GB)

(a) **CLIP model configurations.** EVA CLIP-g can be stably trained via `fp16` precision with fewer image-text pairs (7B *v.s.* 12B / 32B) sampled from a smaller data pool (LAION-400M *v.s.* LAION-2B) on ~1/3×GPUs compared with other open-sourced billion-scale competitors.

datasets	ImageNet-1K [28]	ImageNet-V2 [82]	ImageNet-Adv. [42]	ImageNet-Ren. [41]	ImageNet-Ske. [100]	ObjectNet [6]	CIFAR-10 [54]	CIFAR-100 [54]		UCF-101 [92]	Kinetics-400 [51]	Kinetics-600 [14]	Kinetics-700 [15]	
model	image classification								Δ↓	video classification				avg. all
OpenAI CLIP-L	75.5	69.9	70.8	87.8	59.6	69.0	95.6	75.9	3.4	76.4	64.5	64.2	57.7	72.2
ALIGN	76.4	70.1	75.8	92.2	64.8	72.2	-	-	-	-	-	-	-	-
Open CLIP-H	78.0	70.9	59.3	89.3	66.6	69.7	97.5	84.7	5.7	78.2	63.1	63.6	56.1	73.1
Open CLIP-g	76.6	69.6	57.2	88.7	65.2	67.5	97.1	83.9	5.8	77.7	61.7	62.2	55.0	71.9
EVA CLIP-g	**78.5**(+1.9)	**71.5**(+1.9)	**73.6**(+16.4)	**92.5**(+3.8)	**67.3**(+2.1)	**72.3**(+4.8)	**98.3**(+1.2)	**88.7**(+4.8)	2.5	**76.1**(-1.6)	**65.2**(+3.5)	**64.4**(+2.2)	**58.4**(+3.4)	**75.7**(+3.8)

(b) **Summary of zero-shot image / video classification performance.** "Δ↓": The gap between the averaged performance of ImageNet-{1K, V2, Adv., Ren., Ske.} & ObjectNet that with natural distribution shifts and the original ImageNet-1K validation accuracy. Our model suffers from the smallest performance drop (only 2.5% top-1 accuracy gap) while maintaining the highest zero-shot classification accuracy averaged on all 12 benchmarks (72.7% top-1 accuracy).

Table 10. **EVA as a vision-centric, multi-modal pivot.** We evaluate a billion-scale contrastive language-image pre-trained (CLIP) model with the vision tower initialized from pre-trained EVA, which largely accelerates the contrastive training efficiency and shows promising zero-shot classification performance across a wide range of image / video benchmarks. The statistics & performance of EVA's MIM teacher (OpenAI CLIP-L) are also presented for reference.

limitation (40GB of VRAM): **(i)** relative position biases [90] are not applied. **(ii)** We use 8× decoders in mask2former segmentation head instead of 9×. **(iii)** The feature dimension in mask2former head is ~0.6× of EVA encoder.

Results. We compare EVA with other leading semantic segmentation methods in Table 9. EVA achieves strong results in both ADE20K and COCO-Stuff-164K datasets. On the other hand, the segmentation performance of EVA is slightly lower compared with BEiT-3 on ADE20K, we suspect this is partially due to our weakened architectural configurations.

2.3.5 Contrastive Language-Image Pre-training with Zero-shot Classification Evaluation

CLIP (Contrastive Language-Image Pre-training) [47, 49, 72, 73] is a type of multi-modal foundation model that connects vision and language via contrastive image-text pre-training. CLIP can be applied to any image classification benchmark by simply providing the names of the visual categories to be recognized [1]. Thus the introduction of CLIP essentially reshapes the landscape of visual recognition. Meanwhile, CLIP features also play a central role in representation leaning [70, 101], AI generated content [78, 83, 84] and large dataset filtering [10, 85, 86], *etc*.

In this section and Table 10, we show that EVA is not only a strong encoder for a wide range of vision downstream tasks, but also a multi-modal pivot that builds a bridge between vision and language. To demonstrate that, we train & evaluate EVA as a billion-scale CLIP's vision tower in various zero-shot image / video classification benchmarks.

Baselines and major challenges in CLIP model scaling. We compare our CLIP (dubbed EVA CLIP) with other open-sourced strong CLIP competitors that exploit publicly accessible data / academic resources only. Model configurations and statistics are detailed in Table 10a.

There are two well-known major challenges of CLIP model training and scaling: **(i)** Large-scale Open CLIP models (*e.g*., Open CLIP-H & Open CLIP-g [2, 47]) usually suffer from severe training instability issues [2] and have to use `bfloat16` format for optimization. **(ii)** The training efficiency is low, which may hinder model scaling and downstream performance. For instance, Open CLIP-g is heavily under-trained due to its large compute requirement, and its performance is *even worse* than the sufficiently-trained Open CLIP-H with a smaller model size.

Compared with our CLIP model, Open CLIP-H & -g are trained from scratch with much more image-text pairs (~2.9× and ~1.1× of ours) sampled from a much larger dataset (~5× of ours) on ~3× of GPUs. While by leveraging EVA, billion-scale CLIP model training can be accelerated with improved zero-shot classification performance, described next.

Training settings. For our CLIP model, we initialize the vision encoder via pre-trained EVA and the language encoder

model (SSL)	zero-shot	linear probing	fine-tuning
prev. best	78.0[a]	82.3[b]	89.1[c]
EVA	**78.5**	**86.5**	**89.4**

Table 11. **Zero-shot, linear probing and fine-tuning** performance of **EVA**-CLIP on ImageNet-1K. Notice that the linear probing and fine-tuning results are from the vision encoder of **EVA**-CLIP. Our approach establishes the new state-of-the-art results among all existing self-supervised learning (SSL) methods. (timestamp: Nov 10, 2022) results reference. a: Open CLIP-H [47], b: iBOT [124], c: dBOT [63].

from OpenAI CLIP-L. The pre-training implementation is based on Open CLIP [47]. We also adopt `DeepSpeed` optimization library [80] with ZeRO stage-1 optimizer [77] to save memory. We find using `fp16` format with dynamic loss scaling is stable enough during the whole course of training while using `bfloat16` format is unnecessary. These modifications allow us to train a 1.1B CLIP with a batch size of 41k on 256× NVIDIA A100 40GB GPUs.

Evaluation settings. We evaluate zero-shot image / video classification performance of each CLIP model on 12 benchmarks and report top-1 accuracy for comparisons.

For zero-shot image classification task, we choose 8 benchmarks, *i.e*., ImageNet-1K [28], ImageNet-V2 [81], ImageNet-Adversarial (ImageNet-Adv.) [42], ImageNet-Rendition (ImageNet-Adv.) [41], ImageNet-Sketch (ImageNet-Ske.) [100], ObjectNet [6], CIFAR-10 and CIFAR-100 [54]. We are also interested in the robustness of CLIP models, evaluated via the performance gap between the averaged performance of ImageNet-{1K, V2, Adv., Ren., Ske.} & ObjectNet that with natural distribution shifts and the original ImageNet-1K validation accuracy.

For zero-shot video classification task, we choose 4 benchmarks, namely UCF-101 [92], Kinetics-400 [51], Kinetics-600 [14], and Kinetics-700 [15].

Results. Table 10b shows the comparison. Our **EVA** CLIP achieves the highest averaged accuracy, and performs the best in 10 out of 12 zero-shot classification benchmarks. Notably, the ImageNet-1K validation zero-shot top-1 accuracy is 78.2% without using any of its training set labels, matching the original ResNet-101 [40]. Moreover, our model is quite robust and suffers from the smallest performance drop when facing natural distribution shifts in ImageNet.

At last, in Table 11 we provide zero-shot, linear probing & end-to-end fine-tuning top-1 accuracy of **EVA**-CLIP on ImageNet-1K validation set for reference. Our approach creates the new state-of-the-art results among all existing self-supervised learning methods.

Notice that **EVA** CLIP's vision branch learns from OpenAI CLIP-L, while language branch initialized from the same CLIP-L model. Therefore, starting from a CLIP-L with only 430M parameters, we progressively scale up a 1.1B **EVA** CLIP-g with large performance improvements. This implies that interleaved MIM & image-text contrastive pre-training could be an efficient and scalable CLIP training approach. To our knowledge, **EVA** CLIP-g is the largest performant CLIP model trained via publicly accessible data and resources. We hope our practice on scaling and improving CLIP can also inspire and transfer to the study of other large scale multi-modal foundation models.

3. Related Work

Masked image modeling (MIM) learns rich visual representations via predicting masked visual contents conditioned on visible context. ViT [31] and iGPT [18] report the first meaningful MIM pre-training results. The BEiT family [5,70,101] greatly improves MIM's performance via masked visual token prediction. Recent work [4, 21, 30, 32, 39, 103, 113, 124] (re-)explore pixel / feature regression in MIM, but only in a relatively small model and data scales. In this work, we explore the limits of large scale MIM pre-training via masked image-text aligned feature prediction [43, 106].

Vision foundation models. ConvNets [56] have long been the de-facto standard visual architecture ab initio. Since AlexNet [55], ConvNets have rapidly evolved and become deeper, wider and larger [40,44,66,91,93,94,111]. However, at sufficient model and data scales, ConvNets lag behind ViTs [31] due to a lack of scalable pre-training tasks and the built-in inductive biases. Entering the 2020s, large pre-trained ViTs [31,119] such as SwinV2-G [65] with hierarchical architectures as well as BEiT-3 [101] with multi-modal representations started to demonstrate various vision benchmarks. In this work, we show by leveraging unlabeled images, vanilla ViT can be efficiently scaled up to billion-scale parameters, and stands out in various downstream tasks.

4. Conclusion

In this work, we launch **EVA**, a one billion parameters vanilla ViT encoder to explore the limits of masked visual representation learning. We show simple masked feature modeling as a visual learning pretext task scales well on an architecture with minimal vision priors, and attains excellent results in a representative & diverse set of downstream tasks. We hope **EVA** would bridge the gap between vision and language study via masked modeling, and contributes to the **Neon Genesis** of vision research.

Acknowledgement

This work is supported by the National Key R&D Program of China (2020AAA0105200) and NSFC (No. 61733007). We would like to thank Hanxiao Qu, Yan Tian, Yemin Shi and Xigang Cao for their help on GPU resources. Zhao Xue, Quanyue Ma and Bowen Zhang for their help on datasets and benchmarks, and other colleagues at BAAI for support throughout this project.

References

[1] Clip: Connecting text and images. https://openai.com/blog/clip/. 7

[2] Large scale openclip: L/14, h/14 and g/14 trained on laion-2b. https://laion.ai/blog/large-openclip. 2, 7

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022. 8

[5] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 3, 4, 5, 8

[6] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *NeurIPS*, 2019. 7, 8

[7] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 4

[8] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, 2017. 5

[9] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[10] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 7

[11] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, 2018. 2, 6

[12] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *TPAMI*, 2019. 5, 6

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

[14] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2, 4, 7, 8

[15] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2, 4, 7, 8

[16] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 3

[17] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. 5, 6

[18] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 8

[19] Qiang Chen, Jian Wang, Chuchu Han, Shan Zhang, Zexian Li, Xiaokang Chen, Jiahui Chen, Xiaodi Wang, Shuming Han, Gang Zhang, et al. Group detr v2: Strong object detector with encoder-decoder pretraining. *arXiv preprint arXiv:2211.03594*, 2022. 2, 5, 6

[20] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016. 3

[21] 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. 8

[22] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057*, 2021. 1

[23] 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. 2, 6

[24] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021. 6

[25] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1

[26] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *CVPR*, 2021. 6

[27] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 2021. 4

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 3, 4, 7, 8

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[30] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *arXiv preprint arXiv:2111.12710*, 2021. 8

[31] 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 *ICLR*, 2021. 1, 2, 3, 4, 8

[32] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022. 8

[33] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 6

[34] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 5

[35] WeiFu Fu, CongChong Nie, Ting Sun, Jun Liu, TianLiang Zhang, and Yong Liu. Lvis challenge track technical report 1st place solution: Distribution balanced and boundary refinement for large vo-

cabulary instance segmentation. *arXiv preprint arXiv:2111.02668*, 2021. 2, 6

[36] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D. Cubuk, Quoc V. Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *CVPR*, 2021. 5, 6

[37] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 5, 6

[38] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 2, 5

[39] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 1, 4, 8

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8

[41] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *CVPR*, 2021. 4, 7, 8

[42] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 4, 7, 8

[43] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 1, 3, 8

[44] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 8

[45] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 3

[46] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019. 5

[47] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip. `https://github.com/mlfoundations/open_clip`, 2021. 2, 7, 8

[48] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. *arXiv preprint arXiv:2112.12782*, 2021. 6

[49] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 7

[50] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1

[51] 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. 2, 4, 7, 8

[52] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[53] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *CVPR*, 2020. 6

[54] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7, 8

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 8

[56] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1989. 8

[57] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 2, 6

[58] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 5

[59] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 6

[60] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 5, 6

[61] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *arXiv preprint arXiv:2112.01526*, 2021. 4

[62] 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

[63] Xingbin Liu, Jinghao Zhou, Tao Kong, Xianming Lin, and Rongrong Ji. Exploring target representations for masked autoencoders. *arXiv preprint arXiv:2209.03917*, 2022. 8

[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[65] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 1, 2, 4, 5, 6, 8

[66] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 4, 8

[67] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 3

[68] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022. 5

[69] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019. 3

[70] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1, 2, 3, 4, 5, 7, 8

[71] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*, 2021. 1

[72] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021. 7

[73] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 7

[74] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[75] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021. 1

[76] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 1

[77] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20*, 2020. 3, 8

[78] 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. 7

[79] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser-Nam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *arXiv preprint arXiv:2207.14284*, 2022. 6

[80] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, 2020. 3, 8

[81] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 4, 8

[82] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet?, 2019. 7

[83] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7

[84] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 7

[85] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 7

[86] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 7

[87] 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 *ICCV*, 2019. 3, 5

[88] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3

[89] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 3

[90] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018. 7

[91] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 8

[92] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 7, 8

[93] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 8

[94] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 8

[95] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 5

[96] Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 4

[97] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 3

[98] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *arXiv preprint arXiv:2204.01697*, 2022. 4

[99] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 1

[100] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 4, 7, 8

[101] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[102] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Solo: A simple framework for instance segmentation. *TPAMI*, 2021. 5

[103] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 2, 5, 8

[104] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021. 1

[105] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022. 1, 2

[106] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. *arXiv preprint arXiv:2203.05175*, 2022. 1, 3, 8

[107] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 2, 3, 4, 5, 6

[108] Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019. 4

[109] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Transferring textual knowledge for visual recognition. *arXiv preprint arXiv:2207.01297*, 2022. 2

[110] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, 2020. 4

[111] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 8

[112] 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. 1

[113] 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. *arXiv preprint arXiv:2111.09886*, 2021. 1, 8

[114] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, and Han Hu. On data scaling in masked image modeling. *arXiv preprint arXiv:2206.04664*, 2022. 1

[115] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, 2021. 6

[116] Jianwei Yang, Chunyuan Li, and Jianfeng Gao. Focal modulation networks. *arXiv preprint arXiv:2203.11926*, 2022. 2, 5, 6

[117] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 4, 5

[118] 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. 5, 6

[119] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *CVPR*, 2022. 1, 2, 3, 4, 8

[120] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *arXiv preprint arXiv:2112.07175*, 2021. 5

[121] 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. *arXiv preprint arXiv:2203.03605*, 2022. 5, 6

[122] 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. *arXiv preprint arXiv:2206.05836*, 2022. 6

[123] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 2018. 2, 3, 6

[124] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 1, 2, 8

[125] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Probabilistic two-stage detection. *arXiv preprint arXiv:2103.07461*, 2021. 5