ViTamin: Designing Scalable Vision Models in the Vision-Language Era

Jieneng Chen[1*] Qihang Yu[2*] Xiaohui Shen[2] Alan Yuille[1] Liang-Chieh Chen[2]
[1]Johns Hopkins University [2]ByteDance *equal contribution
https://beckschen.github.io/vitamin.html

Appendix

In the supplementary materials, we provide additional information, as listed below.

- Sec. A: The ablation studies on the ViTamin macro-level network and micro-level block designs.
- Sec. B: ViTamin sets new SoTA in open-vocabulary dense prediction tasks including the OV-LVIS detection benchmark and 6 segmentation benchmarks.
- Sec. C: The results of using the proposed Locked-Text Tuning (LTT) training scheme.
- Sec. D: The results of benchmarking vision models under CLIP setting with an ImageNet-22K data scale.
- Sec. E: The numerical results of benchmarking vision models under CLIP setting with DataComp-1B.
- Sec. F: Detailed results of 38 datasets for different VLMs.
- Sec. G: The training hyper-parameter settings for short/long schedules and high-resolution input fine-tuning.

A. Ablation Studies

We conduct ablation studies on ViTamin design from two aspects: macro-level network and micro-level block. At the macro-level network design, we ablate the hybrid architecture and channel sizes of our three-stage network. At the micro-level block design, we ablate the design choices of convolution blocks and feed-forward network. In the tables, 'IN acc.' and 'avg. 38' denote the ImageNet accuracy (%) and the average accuracy (%) of 38 datasets, respectively. The ImageNet accuracy is used as the main metric. For simplicity, all the ablation studies are performed using base model variants with 128M seen samples.

Hybrid Architecture: In Tab. 1, we ablate design choices of hybrid architectures. Specifically, the compared architectures include ViT-B/16 (pure transformer with TFB or TFB-GeGLU blocks in stage 3), a new MBConvNet-B (pure ConvNet with MBConv-LN blocks in all three stages), and our ViTamin-B (MBConv-LN in stage 1 and 2, and TFB-GeGLU in stage 3). The ablated models may differ in depth but share a similar number of parameters. As shown in the table, our ViTamin-B outperforms both the pure Transformer ViT-B/16 and the pure ConvNet MBConvNet-B by more than +4.7%.

model	# block type		depth	params	IN acc.	avg. 38
	stage 1 & 2	stage 3	stage 3	(M)	(%)	datasets
ViT-B/16	-	TFB	12	86.2	45.8	41.0
ViT-B/16	-	TFB-GeGLU	14	84.2	45.4	40.9
ViT-B/16	-	TFB-GeGLU	15	90.2	46.1	41.2
MBConvNet-B	MBConv-LN	MBConv-LN	18	87.3	45.8	41.7
ViTamin-B	MBConv-LN	TFB-GeGLU	14	87.5	50.8	44.6

Table 1. **Ablation study for hybrid architecture.** MBConv-LN: Mobile Convolution with LayerNorm. TFB-GeGLU: Transformer Block with GeGLU. In this ablation study, we ablate TFB and TFB-GeGLU in ViT-B/16, and design a pure ConvNet using only MBConv-LN across all three stages (called MBConvNet-B in the table). Our final setting is marked in blue.

channel size	params (M)	MACs (G)	IN acc.	avg. 38
(C, 2C, 8C)	86.0	19.5	49.7	44.8
(C, 2C, 6C)	87.5	21.8	50.8	44.6
(C, 2C, 4C)	91.5	28.5	51.0	44.8

Table 2. **Ablation study on the channel sizes.** The channel sizes (x_1C, x_2C, x_3C) denote the channel sizes of stage 1, 2, and 3, respectively, w.r.t. a constant C (*e.g.*, $(x_1, x_2, x_3) = (1, 2, 6)$ and $C = 128$ for ViTamin-B). Our final setting is marked in blue.

block type	params (M)	MACs (G)	IN acc.	avg. 38
ConvNeXt	88.0	21.0	49.8	44.9
MBConv-BN	87.5	21.9	50.5	44.9
MBConv-BN-SE	88.5	21.9	50.9	45.0
MBConv-LN	87.5	21.8	50.8	44.6

Table 3. **Ablation study for design choice of convolutional blocks.** BN: BatchNorm. SE: Squeeze-and-Exciation. LN: LayerNorm. Our final setting is marked in blue.

Channel Sizes of ViTamin: We ablate the effect of varying channel sizes within our ViTamin. The channel sizes (x_1C, x_2C, x_3C) denote the channel sizes of stage 1, 2, and 3, respectively. We set the channel size multipliers x_1 and x_2 to be 1 and 2 (commonly used in the literature for ImageNet). We ablate different values for x_3 in Tab. 2. Our final setting of (C, $2C$, $6C$) improves over (C, $2C$, $8C$) by +1.1%, and is on par with (C, $2C$, $4C$) but uses fewer parameters and MACs.

Design Choice of Convolution Blocks: In Tab. 3, we

image encoder	GeGLU [56]	depth	params (M)	IN acc.	avg. 38
ViT-B/16		12	86.2	45.8	41.0
ViTamin-B		12	89.8	50.3	44.0
ViTamin-B	✓	12	75.7	49.9	43.5
ViTamin-B	✓	14	87.5	50.8	44.6
ViTamin-B		14	104.0	50.4	44.5

Table 4. **Ablation study for design choice of FFN.** Our final setting is marked in blue.

ablate the design choices of convolution blocks in stage 1 and 2. The design choices include ConvNeXt, MBConv-BN, MBConv-BN-SE, and our MBConv-LN. MBConv-BN block is the original MBConv block used in MobileNetv2 [55] with three BatchNorm layers [33], while the MBConv-BN-SE block, proposed by MobileNetv3 [29], augments MBConv-BN with the Squeeze-and-Excitation layer [30]. Each of the MBConv variants demonstrates a superior performance to the ConvNeXt block [43]. Our MBConv-LN, which employs a single Layer Normalization [1], outperforms the MBConv-BN block, and achieves a similar result to MBConv-BN-SE while requiring fewer parameters.

Design Choice of Feed-Forward Network: In Tab. 4, we study the effectiveness of GeGLU [56] in a Transformer Block (TFB) [58]. We experiment with ViT-B/16 and our ViTamin-B, and ablate on the effect of using the original TFB *vs.* the adopted TFB-GeGLU [56]. Remarkably, with the same depth of 12 blocks, ViTamin-B with GeGLU can achieve 49.9% accuracy and surpass the plain ViT-B/16 by a significant +4.1% margin and requires 13% fewer parameters. Adding two more blocks to align the parameters with ViT-B/16, our ViTamin-B boosts its performance to 50.8%, which not only improves over the GeGLU-absent ViTamin-B counterpart (last row) by +0.4% but also maintains a reduced parameters by 26%.

B. Open-Vocabulary Dense Prediction

Frozen Feature Extraction via Sliding Window: We tested the transferability of VLMs to open-vocabulary detection tasks using F-ViT [64] and open-vocabulary segmentation tasks using FC-CLIP [70] frameworks, which both rely on a frozen CLIP backbone. The image size (*e.g.*, 1344×1344) for dense prediction tasks is usually larger than that of upstream VLM pre-training (*e.g.*, 224×224). To employ a frozen transformer-based architecture in these framework, we did not use any distillation [64] or convolutional backbone [70], while we find that a simple sliding window strategy [71] for frozen image feature extraction is effective enough to obtain reasonable performance on downstream tasks requiring high resolution input. The window size is the same as the input image size used during its VLM pretraining. We denote the slightly modified frameworks as *Sliding F-ViT* and *Sliding FC-CLIP*. We follow [64, 70] and use 896×896 and 1344×1344 input size for the open-vocabulary detection and segmentation tasks, respectively.

image encoder	pretraining dataset	pretraining scheme	OV-LVIS [24] (mAP_r)
ViT-L/14	DataComp-1B	CLIPA-v2	32.5
ConvNeXt-L	LAION-2B	OpenCLIP	29.1
ViTamin-L	DataComp-1B	OpenCLIP	**35.6**

Table 5. **Open-vocabulary detection.** Different image encoders (ViT-L/14 by [40] and ConvNeXt-L by [32]) are deployed using the F-ViT framework [64] in a sliding window manner [71], trained on OV-LVIS dataset [24]. ConvNeXt-L is marked in gray due to different pretrained dataset.

B.1. Open-Vocabulary Detection

In Tab.5 of main paper, ViTamin has been validated to be effective for open-vocabulary object detection on the OV-COCO dataset. In this section, we supplement the results on an additional benchmark OV-LVIS, where ViTamin sets a new state-of-the-art performance.

Experimental Setting: The open-vocabulary LVIS (OV-LVIS), introduced in ViLD [24], redefines the 337 rare categories from the LVIS v1.0 [25] dataset as novel categories. We strictly follow the F-ViT [64] framework to perform the open-vocabulary detection tasks, excepting the frozen image features are extracted in a sliding-window manner [71] (denoted as *Sliding F-ViT* in Tab. 6). The effectiveness of VLMs is validated through simply replacing the frozen backbone of F-ViT [64] framework. For evaluation, we follow previous works to use the mean mask AP on rare categories (AP_r) as the metric on OV-LVIS.

Results Analysis: Tab. 5 demonstrates that ViTamin-L is a stronger image encoder for open-vocabulary detector, surpassing its ViT-L/14 counterpart by 3.1% on OV-LVIS dataset [24].

Comparison with Prior Arts: As shown in Tab. 6, ViTamin consistently outperforms all previous methods in the open-vocabulary detection task on OV-LVIS, setting a new state-of-the-art performance of 35.6% AP_r. Notably, our approach surpasses not only the distillation-based backbone (*e.g.*, CLIPSelf [64]) but also larger backbone (*e.g.*, ViT-H/16 in RO-ViT [35]).

B.2. Open-Vocabulary Segmentation

In Tab.6 of main paper, ViTamin has been validated to be effective for open-vocabulary panoptic and semantic segmentation on 8 dataset. We strictly follow the FC-CLIP framework [70] to perform the open-vocabulary segmentation tasks, excepting the frozen image features are extracted in a sliding-window manner [71] (denoted as *Sliding FC-CLIP* in Tab. 5). Following prior works [70], the *Sliding FC-CLIP* is trained on COCO [42] and zero-shot evaluated

detector	image encoder	OV-LVIS (AP_r)	OV-COCO (AP_{50}^{novel})
ViLD [24]	RN50	16.6	27.6
OV-DETR [72]	RN50	17.4	29.4
DetPro [18]	RN50	19.8	-
OC-OVD [3]	RN50	21.1	36.6
OADP [61]	RN50	21.7	-
RegionCLIP [76]	RN50x4	22.0	-
CORA [65]	RN50x4	22.2	41.7
BARON-KD [63]	RN50	22.6	34.0
VLDet [41]	SwinB	26.3	-
F-VLM [38]	RN50x64	32.8	28.0
Detic [78]	SwinB	33.8	-
RO-ViT [35]	ViT-L/16	32.4	33.0
RO-ViT [35]	ViT-H/16	34.1	-
F-ViT [64]	ViT-L/14	24.2	24.7
F-ViT+CLIPSelf [64]	ViT-L/14	34.9	**44.3**
Sliding F-ViT	**ViTamin-L**	**35.6**	37.5

Table 6. **Comparison with prior arts** on open-vocabulary detection on OV-LVIS [24] and OV-COCO [73]. The last row (Sliding F-ViT) shows the result of employing our ViTamin-L using the F-ViT framework [64] in a sliding window manner [71].

method	image encoder	panoptic dataset (PQ)			semantic dataset (mIoU)				
		ADE [77]	Cityscapes [13]	MV [47]	A-150 [77]	A-847 [77]	PC-459 [45]	PC-59 [45]	PAS-21 [19]
FreeSeg [50]	-	16.3	-	-	-	-	-	-	-
OpenSeg [23]	-	-	-	-	21.1	6.3	9.0	42.1	-
GroupViT [67]	ViT-S/16	-	-	-	10.6	6.3	9.0	42.1	-
MaskCLIP [16]	ViT-B/16	15.1	-	-	23.7	8.2	10.0	45.9	-
ODISE [68]	-	22.2	23.9	14.2	29.9	11.1	14.5	57.3	**84.6**
FC-CLIP [70]	ConvNeXt-L	26.8	44.0	**18.3**	34.1	14.8	18.2	58.4	81.8
Sliding FC-CLIP	**ViTamin-L**	**27.3**	**44.0**	18.2	**35.6**	**16.1**	**20.4**	**58.4**	83.4

Table 7. **Comparison with prior arts** on open-vocabulary segmentation. ViTamin sets a new state-of-the-art result on various panoptic and semantic segmenation datasets. The last row (Sliding FC-CLIP) shows the result of employing our ViTamin-L using the FC-CLIP framework [70] in a sliding window manner [71].

on the other datasets. In this section, we compare ViTamin with previous state-of-the-art methods.

Comparison with Prior Arts: In Tab. 7, our approach consistently outperforms all previous open-vocabulary segmentation methods in 2 panoptic dataset and 4 semantic benchmarks, setting a new state-of-the-art. Notably, ViTamin surpasses the the prior art by 0.5% PQ on ADE panoptic dataset and 1.5% mIOU on A-150 semantic dataset.

C. Locked-Text Tuning

Tab. 8 summarizes the detailed results of using the proposed new training scheme, Locked-Text Tuning (LTT). Specifically, when using the LTT training scheme, we employ the text encoder pretrained from ViTamin-L, and use it to guide the training of image encoders of ViTamin-S and ViTamin-B. As shown in the table, we consistently observe the improvements of using LTT. Compared to other distillation-based CLIP training schemes (See the rows marked in grey), our models achieve higher classification and retrieval accuracy in similar model parameters. Practically, despite being adopted from the larger model, the text encoder is much lighter compared to the image encoder (6.6 vs 21.8 GMACs), resulting in only a 14% increase in overall model MACs. Interestingly, using LTT results in a 10% savings in training costs for ViTamin-B, due to the text encoder being fully frozen.

training scheme	training dataset	image encoder	params (M)	seen samp.	IN acc. (%)	avg. 38 (%)	retrieval COCO (%)
models on private/other dataset, for reference							
LiT [75]	Private-4B	ViT-B/32	86.2	0.9B	68.8	-	36.1
TinyCLIP [62]	LAION+YFCC	ViT-45M/32	45.0	1.6B	62.1	-	45.4
TinyCLIP [62]	LAION+YFCC	ViT-63M/32	63.0	1.6B	64.5	-	47.7
our experiments							
OpenCLIP	DataComp-1B	ViTamin-S	22.0	128M	43.3	40.8	25.8
OpenCLIP	DataComp-1B	ViTamin-S	22.0	512M	57.3	49.6	36.6
OpenCLIP	DataComp-1B	ViTamin-S	22.0	1.28B	62.2	53.2	40.2
OpenCLIP	DataComp-1B	ViTamin-B	87.5	128M	50.8	44.6	31.2
OpenCLIP	DataComp-1B	ViTamin-B	87.5	512M	64.0	53.9	41.7
OpenCLIP	DataComp-1B	ViTamin-B	87.5	1.28B	68.9	57.7	44.9
LTT (ours)	DataComp-1B	ViTamin-S	22.0	128M	47.5	44.8	33.4
LTT (ours)	DataComp-1B	ViTamin-S	22.0	512M	58.9	52.0	41.6
LTT (ours)	DataComp-1B	ViTamin-S	22.0	1.28B	63.4	54.6	45.0
LTT (ours)	DataComp-1B	ViTamin-B	87.5	128M	56.7	50.5	39.8
LTT (ours)	DataComp-1B	ViTamin-B	87.5	512M	66.8	57.3	47.1
LTT (ours)	DataComp-1B	ViTamin-B	87.5	1.28B	70.8	59.4	50.0

Table 8. **Locked-Text Tuning (LTT) training scheme.** We use the pretrained text encoder from ViTamin-L and train the image encoders of ViTamin-{S,B}. Due to the use of private or other filtered/merged dataset, the results borrowed from LiT [75] and TinyCLIP [62] are just for reference, and LiT [75] reports retrieval on COCO only. †: a filtered subset of WebLI dataset [9].

D. Benchmarking Vision Models in CLIP with ImageNet-22K Data Scale

Tab. 9 summarizes the results of benchmarking vision models under CLIP setting with ImageNet-22K data scale. Specifically, we mimic the ImageNet-22K data scale by randomly selecting 14.2M data samples from DataComp-1B, and set the training epochs to 90, a standard training setting on ImageNet-22K. Similar to the findings on ImageNet-22K in the literature [43], under such a small data scale (14.2M data samples), ConvNeXt-T consistently outperforms ViT-S/32 and ViT-S/16. However, when the data scales up to 128M, or even 1.28B, the results are totally different, where ViT/16 shows a superior performance to ConvNeXt by a large margin, across all model sizes (see Tab. 10). We note that hybrid models, such as CoAtNet-0 and our ViTamin-S, still demonstrate the best performances under this small data scale, showing that the hybrid design works well across all data sizes.

E. Numerical Results of Benchmarking Vision Models with DataComp-1B

In Fig.2 of the main paper, we provide the analysis of benchmarked results from various aspects. In this section, we further supplement the numerical results of benchmarking vision models (including ViT, ConvNeXt, CoAtNet, and our

image encoder	data size (M)	epoch	#params (M)	MACs (G)	ImageNet Acc.(%)	avg. 38 datasets (%)
ImageNet-22K scale						
ViT-S/32	14.2	90	21.81	1.12	39.4	36.7
ViT-S/16	14.2	90	21.81	4.25	45.7	38.7
ConvNeXt-T	14.2	90	28.61	4.47	45.9	39.3
CoAtNet-0	14.2	90	24.56	4.43	49.1	41.4
ViTamin-S	14.2	90	22.03	5.50	50.3	41.3

Table 9. **Benchmarking vision models under CLIP setting with an ImageNet-22K data scale.** We mimic the ImageNet-22K data scale with 14.2M data size and 90 training epochs (standard training setting on ImageNet-22K). The benchmarked vision models include ViT (pure transformer), ConvNeXt (pure convolution), CoAtNet (hybrid model), and our proposed ViTamin.

ViTamin) across different model scales and data sizes in Tab. 10. As shown in the table, the proposed ViTamin consistently outperforms all the other vision models in almost all settings.

F. Results of 38 dataset for different VLMs.

Tab. 11 demonstrates the detailed results for VLMs with different large-variant image encoders. This table is associated with Tab. 3 of the main paper.

G. Training Hyper-parameter Settings

Tab. 12 and Tab. 13 provide our details of training hyper-parameter settings for short/long schedules and fine-tuning for high resolution, respectively. The short schedule is used to benchmark several vision models on DataComp-1B, along with our ablation studies, while the long schedule is used to train our ViTamin-L for better performances. When fine-tuning the trained model on larger input resolution, we fine-tune with only 200M seen samples and a small constant learning rate.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 6

[3] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *NeurIPS*, 2022. 3

[4] 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. *NeurIPS*, 2019. 6

[5] Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021. 6

[6] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. Winogavil: Gamified association benchmark to challenge vision-and-language models. *NeurIPS*, 2022. 6

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, 2014. 6

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3

[10] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *CVPR*, 2018. 6

[11] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 6

[12] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*, 2011. 6

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 3

[14] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 2021. 5

[15] 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. 6

[16] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023. 3

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5

[18] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 3

[19] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2010. 3, 6

image encoder	image size	text encoder depth / width	seen samples	#params (M) image+text	MACs (G) image+text	ImageNet Acc. (%)	avg. 38 datasets	ImageNet dist. shift.	VTAB	Retrieval
small model variants										
ViT-S/32	224	12 / 384	128M	21.81 + 40.44	1.12 + 1.64	32.1	34.1	25.3	35.8	27.2
ViT-S/16	224	12 / 384	128M	21.81 + 40.44	4.25 + 1.64	38.4	36.7	29.3	38.5	31.3
ConvNeXt-T	224	12 / 384	128M	28.61 + 40.44	4.47 + 1.64	37.0	35.8	30.1	37.5	30.8
CoAtNet-0	224	12 / 384	128M	24.56 + 40.44	4.43 + 1.64	42.4	38.9	33.5	39.9	34.2
ViTamin-S	224	12 / 384	128M	22.03 + 40.44	5.50 + 1.64	**43.3**	**40.8**	**35.6**	**41.0**	**35.2**
ViT-S/32	224	12 / 384	512M	21.81 + 40.44	1.12 + 1.64	47.3	44.3	36.7	46.9	36.9
ViT-S/16	224	12 / 384	512M	21.81 + 40.44	4.25 + 1.64	53.8	46.5	41.9	46.7	42.1
ConvNeXt-T	224	12 / 384	512M	28.61 + 40.44	4.47 + 1.64	53.3	46.1	42.4	46.4	42.2
CoAtNet-0	224	12 / 384	512M	24.56 + 40.44	4.43 + 1.64	56.4	49.0	45.2	**49.0**	45.0
ViTamin-S	224	12 / 384	512M	22.03 + 40.44	5.50 + 1.64	**57.3**	**49.6**	**46.9**	48.8	**45.4**
ViT-S/32	224	12 / 384	1.28B	21.81 + 40.44	1.12 + 1.64	53.0	47.3	41.3	48.9	41.5
ViT-S/16	224	12 / 384	1.28B	21.81 + 40.44	4.25 + 1.64	59.8	50.9	47.2	51.3	47.5
ConvNeXt-T	224	12 / 384	1.28B	28.61 + 40.44	4.47 + 1.64	59.9	51.3	47.8	52.7	48.3
CoAtNet-0	224	12 / 384	1.28B	24.56 + 40.44	4.43 + 1.64	61.7	51.6	50.1	51.3	48.9
ViTamin-S	224	12 / 384	1.28B	22.03 + 40.44	5.50 + 1.64	**62.2**	**53.2**	**51.3**	**51.7**	**50.0**
base model variants										
ViT-B/32	224	12 / 512	128M	86.19 + 63.43	4.37 + 2.91	38.9	38.0	30.6	40.6	30.7
ViT-B/16	224	12 / 512	128M	86.19 + 63.43	16.87 + 2.91	45.8	41.0	35.8	42.1	36.2
ConvNeXt-B	224	12 / 512	128M	88.09 + 63.43	15.38 + 2.91	41.4	39.7	33.5	41.2	34.1
CoAtNet-2	224	12 / 512	128M	74.18 + 63.43	15.94 + 2.91	48.5	43.5	38.9	43.8	39.1
ViTamin-B	224	12 / 512	128M	87.53 + 63.43	21.84 + 2.91	**50.8**	**44.6**	**41.3**	**45.1**	**40.8**
ViT-B/32	224	12 / 512	512M	86.19 + 63.43	4.37 + 2.91	54.8	48.3	42.7	50.1	42.4
ViT-B/16	224	12 / 512	512M	86.19 + 63.43	16.87 + 2.91	60.0	51.0	48.2	51.4	47.5
ConvNeXt-B	224	12 / 512	512M	88.09 + 63.43	15.38 + 2.91	59.4	50.3	47.9	49.9	47.2
CoAtNet-2	224	12 / 512	512M	74.18 + 63.43	15.94 + 2.91	63.3	52.4	52.4	51.0	49.7
ViTamin-B	224	12 / 512	512M	87.53 + 63.43	21.84 + 2.91	**64.0**	**53.9**	**53.3**	**53.7**	**50.8**
ViT-B/32	224	12 / 512	1.28B	86.19 + 63.43	4.37 + 2.91	60.1	52.5	47.4	53.6	47.5
ViT-B/16	224	12 / 512	1.28B	86.19 + 63.43	16.87 + 2.91	65.6	55.6	53.1	55.3	51.7
ConvNeXt-B	224	12 / 512	1.28B	88.09 + 63.43	15.38 + 2.91	65.3	54.7	54.0	54.2	51.7
CoAtNet-2	224	12 / 512	1.28B	74.18 + 63.43	15.94 + 2.91	68.5	56.8	57.2	56.0	53.4
ViTamin-B	224	12 / 512	1.28B	87.53 + 63.43	21.84 + 2.91	**68.9**	**57.7**	**58.3**	**56.4**	**54.1**
large model variants										
ViT-L/32	224	12 / 768	128M	303.97 + 123.65	15.27 + 6.55	43.5	40.8	34.0	42.7	34.2
ViT-L/16	224	12 / 768	128M	303.97 + 123.65	59.70 + 6.55	49.4	43.8	38.7	44.3	38.9
ViT-L/14	224	12 / 768	128M	303.97 + 123.65	77.83 + 6.55	49.9	43.8	39.4	44.5	39.3
ConvNeXt-XL	224	12 / 768	128M	350.25 + 123.65	79.65 + 6.55	42.8	38.4	33.3	38.4	35.0
CoAtNet-4	224	12 / 768	128M	275.07 + 123.65	60.81 + 6.55	52.5	**45.2**	42.0	**45.2**	41.1
ViTamin-L	224	12 / 768	128M	333.32 + 123.65	72.60 + 6.55	**52.7**	44.8	**42.4**	44.6	**41.8**
ViT-L/32	224	12 / 768	512M	303.97 + 123.65	15.27 + 6.55	60.4	51.8	47.4	52.7	47.3
ViT-L/16	224	12 / 768	512M	303.97 + 123.65	59.70 + 6.55	66.4	55.6	53.6	55.5	52.2
ViT-L/14	224	12 / 768	512M	303.97 + 123.65	77.83 + 6.55	67.0	55.4	54.8	54.2	52.0
ConvNeXt-XL	224	12 / 768	512M	350.25 + 123.65	79.65 + 6.55	63.0	52.5	51.1	51.8	49.4
CoAtNet-4	224	12 / 768	512M	275.07 + 123.65	60.81 + 6.55	66.8	56.1	56.4	**56.5**	50.4
ViTamin-L	224	12 / 768	512M	333.32 + 123.65	72.60 + 6.55	**68.7**	**56.6**	**56.8**	**56.5**	**53.2**
ViT-L/32	224	12 / 768	1.28B	303.97 + 123.65	15.27 + 6.55	67.5	57.0	54.1	57.9	51.9
ViT-L/16	224	12 / 768	1.28B	303.97 + 123.65	59.70 + 6.55	71.9	60.1	59.9	59.9	56.0
ViT-L/14	224	12 / 768	1.28B	303.97 + 123.65	77.83 + 6.55	72.3	60.7	60.5	60.0	56.0
ConvNeXt-XL	224	12 / 768	1.28B	350.25 + 123.65	79.65 + 6.55	70.2	58.3	59.1	57.0	55.5
CoAtNet-4	224	12 / 768	1.28B	275.07 + 123.65	60.81 + 6.55	71.3	59.4	61.4	59.1	53.4
ViTamin-L	224	12 / 768	1.28B	333.32 + 123.65	72.60 + 6.55	**73.9**	**62.0**	**62.9**	**61.4**	**56.6**

Table 10. **Benchmarking vision backbones on Datacomp-1B under CLIP setting (contrastive language-image pretraining).** We benchmark popular vision backbones, including ViT [17] (pure transformer model), ConvNeXt [43] (pure convolution model), CoAtNet [14] (hybrid convolution-transformer model), and our proposed ViTamin, under different model parameters and training seen samples.

image encoder	training scheme	avg. 38	ImageNet 1k [15]	Caltech-101 [20]	CIFAR-10 [37]	CIFAR-100 [37]	CLEVR Counts [34]	CLEVR Distance [37]	Country211 [51]	Describable Textures [11]	EuroSAT [26]	FGVC Aircraft [44]	Food-101 [7]	GTSRB [57]	ImageNet Sketch [60]	ImageNet v2 [53]	ImageNet-A [28]	ImageNet-O [28]	ImageNet-R [27]
ViT-L/14	[32]	66.3	79.2	94.7	98.2	87.3	35.6	24.4	31.6	66.5	71.2	47.5	94.5	58.5	68.0	72.1	69.6	32.6	90.8
ViT-L/14	[40]	65.4	79.6	94.5	98.7	88.5	18.6	24.5	29.4	69.6	60.4	43.0	94.2	59.1	70.6	73.0	71.2	33.7	92.9
ViT-L/14 †	[40]	65.7	80.3	94.4	98.6	88.3	15.7	24.4	30.7	68.6	58.1	42.8	94.6	57.0	70.9	73.5	77.7	32.9	93.3
ViTamin-L	[32]	66.7	80.8	95.1	98.5	88.0	35.4	24.1	33.0	68.1	66.6	49.2	94.9	61.9	71.5	73.6	72.4	32.0	93.0
ViTamin-L†	[32]	67.2	81.8	95.6	98.5	87.8	31.7	24.1	36.0	69.2	64.7	49.7	95.8	63.4	72.1	75.2	81.7	31.1	93.8
ViTamin-XL†	[32]	**68.1**	82.6	95.7	98.7	88.8	19.1	20.0	37.8	71.5	75.6	53.7	96.0	53.2	73.1	76.3	83.1	33.0	94.2

image encoder	training scheme	KITTI Vehicle Distance [22]	MNIST [39]	ObjectNet [4]	Oxford Flowers-102 [48]	Oxford-IIIT Pet [49]	Pascal VOC 2007 [19]	PatchCamelyon [59]	Rendered SST2 [74]	RESISC45 [74]	Stanford Cars [36]	STL-10 [12]	SUN397 [66]	SVHN [46]	Flickr [69]	MSCOCO [8]	WinoGAViL [6]	iWildCam [5]	Camelyon17 [2]	FMoW [10]	Dollar Street [54]	GeoDE [52]
ViT-L/14	[32]	27.9	86.6	74.3	82.6	95.1	82.5	51.2	61.0	69.4	93.1	99.3	74.3	67.7	81.2	54.5	46.7	16.1	50.9	24.0	66.2	91.5
ViT-L/14	[40]	19.3	73.7	69.9	81.0	95.0	80.7	59.2	53.9	68.4	93.7	99.2	75.3	63.9	81.9	56.0	43.9	17.2	67.6	24.6	66.5	91.5
ViT-L/14 †	[40]	20.0	76.7	73.2	81.0	95.0	79.8	60.3	53.2	68.8	94.1	99.3	75.6	62.9	82.5	56.4	44.5	19.4	67.8	25.0	67.5	92.4
ViTamin-L	[32]	23.6	81.8	76.1	85.0	95.6	80.2	52.2	61.1	73.0	94.8	99.3	75.8	66.1	81.1	54.7	45.1	17.2	58.0	16.4	68.1	91.5
ViTamin-L†	[32]	21.2	81.3	80.7	84.7	95.8	82.0	50.1	60.5	73.1	95.1	99.5	76.3	66.9	82.5	55.7	47.3	19.1	49.5	17.5	70.7	92.2
ViTamin-XL†	[32]	17.4	88.9	81.9	85.6	95.9	83.8	56.2	61.9	75.9	94.8	99.4	76.2	74.0	84.7	58.7	47.9	21.4	46.0	22.5	68.6	92.5

Table 11. **Detailed results of 38 dataset for different VLMs.** The compared models are trained with the scheme of either OpenCLIP [32] or CLIPA-v2 [40]. All models are trained on DataComp-1B [21] dataset with similar seen samples for a fair comparison. †: using larger number of patches of 576 (*i.e*., image size of 336 for row 3 and 384 for row 5, respectively).

training config	*short schedule* ViTamin-S/B/L 224^2	*long schedule* ViTamin-L/L2/XL/XL $224^2/224^2/256^2/256^2$
batch size	8k/8k/16k	90k
seen samples	1.28B	12.8B/12.8B/12.8B/40B
optimizer	AdamW	AdamW
base learning rate	5e-4	2e-3
weight decay	0.02	0.02
optimizer momentum β_1	0.9	0.9
optimizer momentum β_2	0.98/0.98/0.95	0.95
learning rate schedule	cosine decay	cosine decay
warmup steps	500	782/4436/4436/9981
warmup schedule	linear	linear
random crop ratio	none	[0.4, 1.0]
stochastic depth [31]	0.1	0.1
precision	amp bfloat16	amp bfloat16

Table 12. **Short/Long schedule training settings for ViTamin variants**.

pre-training config	ViTamin-L 224^2	ViTamin-L2 224^2	ViTamin-XL 256^2
fine-tuning config	$256^2/336^2/384^2$	$256^2/336^2/384^2$	$256^2/384^2$
batch size	90k	90k	90k
seen samples	0.2B	0.5B	0.5B
optimizer	AdamW	AdamW	AdamW
base learning rate	1e-5	1e-5	1e-5
weight decay	0	0	0
optimizer momentum β_1	0.9	0.9	0.9
optimizer momentum β_2	0.95	0.95	0.95
learning rate schedule	constant	constant	constant
warmup steps	0	0	0
random crop ratio	none	none	none
stochastic depth [31]	0.1	0.1	0.1
precision	amp bfloat16	amp bfloat16	amp bfloat16

Table 13. **Fine-tuning setting for high resolution**. The models are pre-trained with *long schedule* and then fine-tuned on the target resolution.

[20] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPRW*, 2004. 6

[21] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 6

[22] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 6

[23] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 3

[24] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 3

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

[26] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 6

[27] 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 *ICCV*, 2021. 6

[28] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 6

[29] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019. 2

[30] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2

[31] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 6

[32] 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, 2021. 2, 6

[33] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 2

[34] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 6

[35] Dahun Kim, Anelia Angelova, and Weicheng Kuo. Region-aware pretraining for open-vocabulary object detection with vision transformers. In *CVPR*, 2023. 2, 3

[36] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, 2013. 6

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[38] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2023. 3

[39] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 6

[40] Xianhang Li, Zeyu Wang, and Cihang Xie. Scaling clip training with 81.1% zero-shot imagenet accuracy within a $10,000 budget; an extra $4,000 unlocks 81.8% accuracy. *arXiv preprint arXiv:2306.15658*, 2023. 2, 6

[41] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. *arXiv preprint arXiv:2211.14843*, 2022. 3

[42] 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 *ECCV*, 2014. 2

[43] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2, 3, 5

[44] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[45] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 3

[46] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[47] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 3

[48] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 6

[49] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 6

[50] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 2023. 3

[51] 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. 6

[52] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron B Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *arXiv preprint arXiv:2301.02560*, 2023. 6

[53] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 6

[54] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. *NeurIPS*, 2022. 6

[55] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, 2018. 2

[56] Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020. 2

[57] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, 2011. 6

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 2

[59] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, 2018. 6

[60] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 6

[61] Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, and Si Liu. Object-aware distillation pyramid for open-vocabulary object detection. In *CVPR*, 2023. 3

[62] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*, 2023. 3

[63] Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, and Chen Change Loy. Aligning bag of regions for open-vocabulary object detection. In *CVPR*, 2023. 3

[64] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2, 3

[65] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, 2023. 3

[66] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 119:3–22, 2016. 6

[67] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, 2022. 3

[68] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 3

[69] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 6

[70] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *NeurIPS*, 2023. 2, 3

[71] Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. Towards open-ended visual recognition with large language model. *arXiv preprint arXiv:2311.08400*, 2023. 2, 3

[72] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022. 3

[73] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3

[74] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. The visual task adaptation benchmark. 2019. 6

[75] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 3

[76] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 3

[77] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 3

[78] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 3