Seeing What Matters: Empowering CLIP with Patch Generation-to-Selection

Supplementary Material

Appendix

This appendix presents implementation details (§A), supplementary experiments (§B), and additional qualitative results (§C).

A. Implementation Details

A.1. Architecture

Consistent with the CLIP [25] and OpenCLIP [4] frameworks, we utilize the ViT-B/16 [8] architecture as the primary image encoder backbone, paired with a 12-layer transformer featuring 512-dimensional embeddings and 8 attention heads for the text encoder. This appendix also includes experiments with ViT-S/16, a smaller vision backbone with 384-dimensional embeddings, 12 layers, and 6 attention heads. Table A1 summarizes the detailed configurations of both ViT-S/16 and ViT-B/16, including their parameter counts and architectural specifications for both the vision and text encoders.

A.2. Pre-training

We provide the detailed pre-training settings for CLIP-PGS in Table A2. The model is optimized using AdamW [20] with momentum parameters set to $(0.9, 0.98)$, a learning rate of 1×10^{-3}, and a weight decay of 0.2. A cosine decay schedule [19] with 10,000 warm-up steps is applied. The training is conducted for 32 epochs with a batch size of 4,096, utilizing an environment of 8 NVIDIA V100 GPUs (32G). These settings ensure efficient optimization and scalability during pre-training.

Model	Embed Dim.	Vision Transformer			Text Transformer			# Params (M)		
		Layers	Width	Heads	Layers	Width	Heads	Vision	Text	Total
S/16	384	12	384	6	12	384	6	22	33	55
B/16	512	12	768	12	12	512	8	86	53	141

Table A1. **Detailed configuration of encoder architectures**, including embedding dimensions (Dim.), transformer specifications for vision and text encoders, and parameter counts.

Config	Value
optimizer	AdamW [20]
optimizer momentum	(0.9, 0.98)
batch size	4,096
learning rate	1e-3
warm-up steps	10,000
schedule	cosine decay [19]
weight decay	0.2
training epochs	32
GPU environment	8 NVIDIA V100 GPUs (32G)

Table A2. Pre-training settings of CLIP-PGS.

Dataset	Classes	Train Size	Test Size	Evaluation Task
Food101 [1]	101	75,750	25,250	fine-grained recognition
CIFAR10 [16]	10	45,000	10,000	fine-grained recognition
CIFAR100 [16]	100	45,000	10,000	fine-grained recognition
SUN397 [32]	397	-	108,754	scene recognition
Cars [15]	196	8,144	8,041	fine-grained recognition
VOC2007 [9]	20	7,844	14,976	object recognition
Aircraft [21]	100	3,334	3,333	fine-grained recognition
DTD [5]	47	1,880	1,880	texture recognition
OxfordPets [24]	37	2,944	3,669	fine-grained recognition
Caltech101 [10]	102	2,753	6,085	object recognition
Flowers [23]	102	1,020	6,149	fine-grained recognition
STL10 [6]	10	5,000	8,000	object recognition
EuroSAT [11]	10	16,200	5,400	aerial image recognition
RESISC45 [3]	45	18,900	6,300	aerial image recognition
GTSRB [27]	43	26,640	12,630	traffic sign recognition
Country211 [28]	211	31,650	21,100	geo-tagged recognition
PCam [29]	2	262,144	32,768	digital pathology
ImageNet-1K [7]	1000	1,281,167	50,000	fine-grained recognition
ImageNet-V2 [26]	1000	-	10,000	robustness of collocation
ImageNet-A [13]	200	-	7,500	robustness of attack
ImageNet-R [12]	200	-	30,000	robustness of multi-domains
ImageNet-O [13]	200	-	7,500	robustness of attack
ImageNet-Sketch [30]	1000	-	50,889	robustness of sketch domain
MS-COCO [18]	-	82,783	5,000	text/image retrieval
Flickr8K [34]	-	6,000	1,000	text/image retrieval
Flickr30K [34]	-	29,000	1,000	text/image retrieval

Table B3. **Overview of downstream datasets**, including the number of classes, training and testing set sizes, and evaluation tasks.

B. Supplementary Experiments

B.1. Downstream Datasets.

Table B3 provides a comprehensive overview of the datasets utilized in our experiments, detailing the number of classes, training and testing set sizes, and their corresponding evaluation tasks (*e.g.*, recognition, robustness, and retrieval).

B.2. Downstream Evaluation Tasks

In this supplementary material, we expand the evaluation to include results with ViT-S/16 [8] as the visual backbone, complementing the main text. We assess the model across **five** standard benchmark scenarios: zero-shot classification, zero-shot text/image retrieval, linear probing, robustness evaluation, and language compositionality, adhering to established evaluation protocols [4, 17, 22, 25]. [1]

- **Zero-shot classification** (Table B4): We evaluate model generalizability on 17 datasets, including Food101 [1], CIFAR10 [16], and ImageNet-1K [7]. These benchmarks assess performance under varying distributional shifts, highlighting the robustness of our approach.
- **Zero-shot retrieval** (Table B5): Text-to-image and image-to-text retrieval tasks are conducted on MS-

[1] https://github.com/LAION-AI/CLIP_benchmark

COCO [18] and Flickr [34]. These benchmarks evaluate the model's capability to associate visual and language representations without additional fine-tuning.

- **Linear probing** (Table B6): Visual representations are assessed on ImageNet-1K [7], CIFAR10 [16], and CIFAR100 [16] using linear classifiers. Following the clip-benchmark setup [1], we train for 10 epochs with AdamW optimizer [20], a 0.1 learning rate, and a batch size of 64.
- **Robustness evaluation** (Table B7): The model is tested on ImageNet-1K [7] and out-of-distribution datasets such as ImageNet-V2 [26], ImageNet-A [13], and ImageNet-Sketch [30]. This evaluation examines resilience to distributional shifts.
- **Language compositionality** (Table B8): Performance on the SugarCrepe [14] dataset is used to assess adaptability to complex language structures, including manipulations of objects, attributes, and relations, showcasing the model's precision in aligning visual and linguistic cues.

B.3. Ablation Studies

This section presents an in-depth ablation analysis of the key design components in CLIP-PGS. Unless otherwise specified, experiments use ViT-B/16 as the image encoder, trained on the CC12M dataset [2] for 32 epochs with a batch size of 4,096. We evaluate performance on diverse downstream tasks, including zero-shot classification (ZS), linear probing (LP), and zero-shot text/image retrieval (TR/IR), as summarized in Table B9. For edge detection (ED), we compare the commonly used Sobel operator with the Canny edge detector to investigate their impact on performance.

C. Additional Qualitative Results

Classification Results. This section highlights the qualitative performance of CLIP-PGS in zero-shot classification and robustness tasks, showcasing its adaptability and generalization. For zero-shot classification (Fig. C1), examples from diverse datasets demonstrate the model's accurate alignment of visual and textual representations, adapting effectively to varied categories and contexts. For zero-shot robustness (Fig. C2), results from robustness-focused datasets illustrate CLIP-PGS's resilience in handling distributional shifts while maintaining high prediction quality.

Retrieval Results. This section evaluates CLIP-PGS on zero-shot text and image retrieval tasks using the MS-COCO [18] dataset. For text retrieval (Fig. C3), the model retrieves highly relevant images for given text queries, showcasing precise alignment between textual and visual content. For image retrieval (Fig. C4), CLIP-PGS effectively links input images to their corresponding textual descriptions, demonstrating robust cross-modal associations even in complex scenarios.

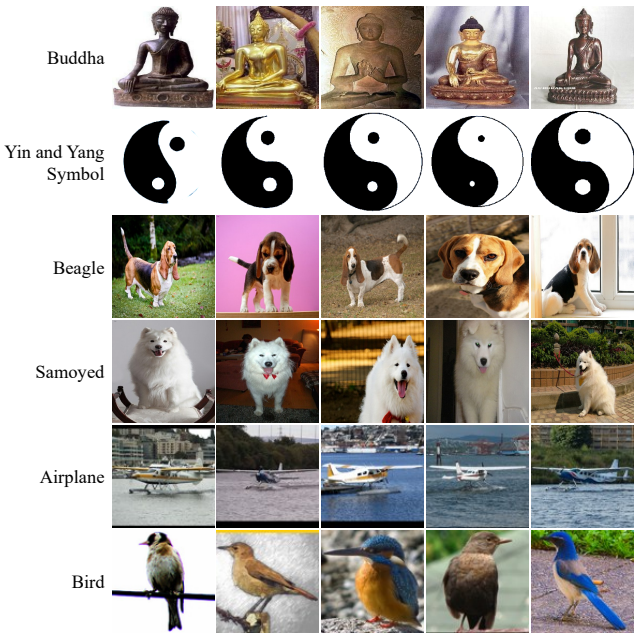

Figure C1. **Visualization of zero-shot classification results**. We provide the top-5 predictions of our proposed CLIP-PGS$_{0.3}$. The first two rows report examples from Caltech101 [10], the next two rows highlight samples from OxfordPets [24], and the final two rows present results from STL10 [6].

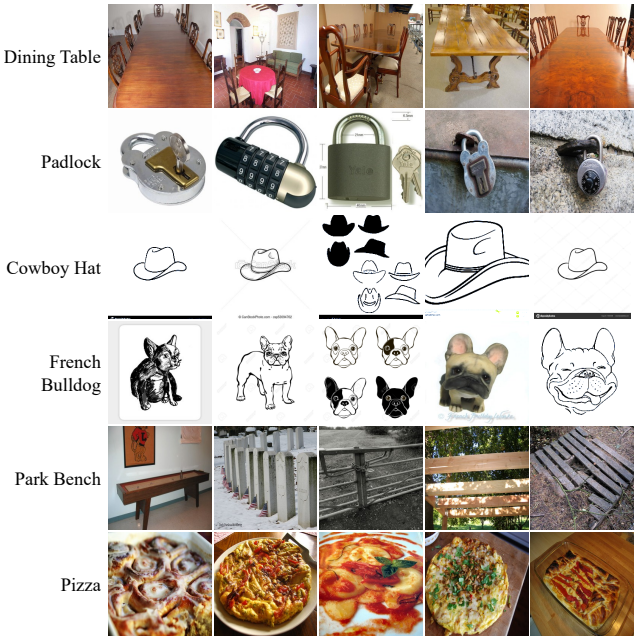

Figure C2. **Visualization of zero-shot robustness results**. We provide the top-5 predictions of our proposed CLIP-PGS$_{0.3}$. The first two rows report examples from ImageNet-1K [7], the next two rows highlight samples from ImageNet-R [12], and the final two rows present results from ImageNet-O [13].

Method	Image Enc.	Food101	CIFAR10	CIFAR100	SUN397	Cars	VOC2007	Aircraft	DTD	OxfordPets	Caltech101	Flowers	STL10	EuroSAT	RESISC45	GTSRB	Country211	PCam	Average
CLIP [25]	ViT-B/16	42.3	57.7	25.0	44.1	17.0	50.5	1.7	16.5	53.9	73.5	26.0	82.0	18.7	26.5	9.4	4.5	48.0	35.1
FLIP [17]	ViT-B/16	39.9	52.8	24.5	42.8	15.9	46.6	1.4	15.9	46.0	70.4	25.3	80.2	17.0	25.8	5.6	4.0	47.1	33.0
A-CLIP [33]	ViT-B/16	41.8	61.6	27.1	46.6	16.0	51.1	1.3	17.1	51.2	73.5	25.7	85.8	20.5	29.1	8.0	4.2	50.1	35.9
E-CLIP [31]	ViT-B/16	42.1	70.7	32.0	43.9	15.1	43.6	2.2	17.0	55.4	**73.7**	28.4	85.6	22.9	30.0	9.6	**4.7**	50.0	36.9
Ours																			
CLIP-PGS$_{0.5}$	ViT-B/16	42.8	62.5	35.5	45.5	17.3	50.0	1.9	17.4	55.7	71.8	**33.2**	88.2	20.5	**31.8**	10.1	**4.7**	50.0	37.6
CLIP-PGS$_{0.3}$	ViT-B/16	**46.5**	**73.5**	**37.3**	**47.5**	**19.9**	**55.1**	**3.1**	**19.8**	**58.1**	72.7	30.7	88.2	22.8	30.4	10.9	4.5	50.8	**39.5**
CLIP-PGS$_{0.5}$	ViT-S/16	38.7	58.4	29.0	43.7	12.7	48.0	2.0	17.7	50.6	69.4	26.6	86.4	**27.6**	25.9	**11.2**	3.9	56.4	35.8
CLIP-PGS$_{0.3}$	ViT-S/16	39.1	66.9	30.8	44.0	15.0	47.1	2.7	14.8	54.5	71.8	28.5	**88.3**	16.6	25.6	7.9	4.3	50.2	35.8

Table B4. **Zero-shot classification results**. We evaluate performance on 17 diverse classification datasets, reporting both top-1 accuracy (%) and the overall average. The optimal result is highlighted in **bold**, and the second-best result is underlined.

Method	Image Enc.	Text Retrieval									Image Retrieval								
		MS-COCO			Flickr8K			Flickr30K			MS-COCO			Flickr8K			Flickr30K		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLIP [25]	ViT-B/16	34.6	62.0	72.7	55.7	81.6	89.9	58.5	83.8	89.1	23.5	47.8	59.7	40.5	68.9	80.2	43.2	70.4	80.4
FLIP [17]	ViT-B/16	32.6	59.1	70.6	55.0	80.9	88.9	53.8	80.8	88.5	22.6	46.1	58.1	40.3	68.1	78.6	41.5	67.9	77.5
A-CLIP [33]	ViT-B/16	33.7	60.2	71.0	53.7	80.1	88.0	55.3	81.4	87.6	23.9	48.3	60.0	40.6	68.9	78.9	43.1	70.1	78.8
E-CLIP [31]	ViT-B/16	34.3	62.0	73.3	57.0	82.7	90.1	55.8	**84.2**	89.6	23.8	48.2	59.8	42.0	69.4	79.6	43.3	70.9	80.2
Ours																			
CLIP-PGS$_{0.5}$	ViT-B/16	35.2	61.9	72.8	**58.5**	**83.6**	90.6	57.7	82.7	90.4	24.3	48.8	60.5	43.5	70.7	81.0	45.3	72.9	81.2
CLIP-PGS$_{0.3}$	ViT-B/16	**36.0**	**64.4**	**74.6**	58.3	82.9	**90.8**	**59.9**	83.5	**90.8**	**25.1**	**49.5**	**61.6**	**44.4**	**71.7**	**81.1**	**47.1**	**73.5**	**82.0**
CLIP-PGS$_{0.5}$	ViT-S/16	31.6	58.1	69.6	53.6	81.5	89.3	53.9	80.0	87.4	22.2	45.6	57.6	39.8	68.1	78.8	41.4	68.2	77.3
CLIP-PGS$_{0.3}$	ViT-S/16	33.1	60.5	72.4	52.3	80.9	89.7	55.9	80.5	88.1	22.9	46.7	58.6	40.2	68.1	79.4	42.7	69.1	78.0

Table B5. **Zero-shot text/image retrieval results**. We evaluate performance on the MS-COCO [18], Flickr8k [34], and Flickr30k [34] datasets, reporting Recall@1 (%, R@1), Recall@5 (%, R@5), and Recall@10 (%, R@10) for both text and image retrieval tasks.

Method	Image Enc.	CIFAR10	CIFAR100	ImageNet-1K
CLIP [25]	ViT-B/16	88.0	67.4	62.3
FLIP [17]	ViT-B/16	85.9	65.5	61.3
A-CLIP [33]	ViT-B/16	86.4	66.1	62.0
E-CLIP [31]	ViT-B/16	89.0	69.7	62.7
Ours				
CLIP-PGS$_{0.5}$	ViT-B/16	89.5 **(+0.5)**	70.3 **(+0.6)**	64.2 **(+1.5)**
CLIP-PGS$_{0.3}$	ViT-B/16	**90.0 (+1.0)**	**72.3 (+2.6)**	**64.4 (+1.7)**
CLIP-PGS$_{0.5}$	ViT-S/16	87.7	68.6	62.7
CLIP-PGS$_{0.3}$	ViT-S/16	88.1	68.7	62.9

Table B6. **Linear probing classification results**. We evaluate all models on three common datasets, *i.e.*, CIFAR10 [16], CIFAR100 [16], and ImageNet-1K [7], training each for 10 epochs under a consistent linear training setup. We present top-1 accuracy (%), with gains over the stronger baseline highlighted in **(green)**.

Figure C3. **Visualization of zero-shot text retrieval**. We provide the top-3 predictions of our proposed CLIP-PGS$_{0.3}$. Examples are from the retrieval dataset MS-COCO [18].

Method	Image Enc.	ImageNet-1K	ImageNet-V2	ImageNet-A	ImageNet-R	ImageNet-O	ImageNet-Sketch	Average	ID Average	OOD Average
CLIP [25]	ViT-B/16	36.1	30.7	8.0	47.6	38.4	24.9	31.0	36.1	29.0
FLIP [17]	ViT-B/16	34.4	29.5	7.1	41.4	39.5	20.1	28.7	34.4	27.5
A-CLIP [33]	ViT-B/16	35.2	30.1	8.1	45.1	39.4	23.7	30.3	35.2	30.3
E-CLIP [31]	ViT-B/16	36.3	30.7	8.1	47.9	39.6	25.4	31.3	36.3	30.3
Ours										
CLIP-PGS$_{0.5}$	ViT-B/16	38.0	32.6	9.1	45.1	41.1	23.9	31.6	38.0	30.4
CLIP-PGS$_{0.3}$	ViT-B/16	**38.6**	**33.1**	**9.6**	**48.1**	**42.6**	**25.6**	**32.9**	**38.6**	**31.8**
CLIP-PGS$_{0.5}$	ViT-S/16	34.9	29.5	7.4	41.8	40.6	21.4	29.3	34.9	28.1
CLIP-PGS$_{0.3}$	ViT-S/16	35.4	30.2	7.8	43.5	40.4	22.5	30.0	35.4	28.9

Table B7. **Robustness assessment results**. We evaluate model robustness on ImageNet-1K [7] and five of its variants [12, 13, 26, 30], reporting top-1 accuracy (%) along with overall averages for in-distribution (ID) and out-of-distribution (OOD) performance.

Method	Image Enc.	REPLACE			SWAP		ADD		Average		
		Object	Attribute	Relation	Object	Attribute	Object	Attribute	Object	Attribute	Relation
CLIP [25]	ViT-B/16	85.8	**79.2**	64.5	61.8	58.7	74.2	68.4	73.7	68.8	64.5
FLIP [17]	ViT-B/16	84.1	75.9	66.0	60.2	61.6	71.7	63.2	72.0	66.9	66.0
A-CLIP [33]	ViT-B/16	86.6	75.5	63.2	52.4	63.1	71.6	66.8	71.6	68.4	63.2
E-CLIP [31]	ViT-B/16	86.9	73.5	60.2	59.4	63.4	73.3	66.8	73.2	68.4	60.2
Ours											
CLIP-PGS$_{0.5}$	ViT-B/16	86.0	77.0	64.6	63.3	65.5	**77.3**	69.8	**75.5**	**70.8**	64.6
CLIP-PGS$_{0.3}$	ViT-B/16	**88.1**	76.0	**67.9**	**64.1**	**66.5**	74.2	**69.9**	**75.5**	**70.8**	**67.9**
CLIP-PGS$_{0.5}$	ViT-S/16	84.9	76.5	65.4	60.4	65.0	73.6	**69.9**	73.0	70.5	65.4
CLIP-PGS$_{0.3}$	ViT-S/16	86.6	77.8	63.9	58.8	63.4	74.9	68.4	73.4	70.2	63.9

Table B8. **Language compositionality results**. We evaluate the compositionality of vision-language models on the SugarCrepe [14] dataset, which tests models by generating mismatched captions by replacing, swapping, or adding fine-grained atomic concepts (object, attribute, and relation). We report Recall@1 (%) and the overall average for each atomic concept.

Method	Component			ImageNet-1K		MS-COCO	
	MR	ED	OTN	ZS	LP	TR	IR
Baseline CLIP [25]	-	-	-	36.1	62.3	34.6	23.5
Random Mask FLIP [17]	0.5	-	-	34.4	61.3	32.6	22.6
CLIP-PGS$_{0.5}$	0.5	✗	✗	35.2	61.9	33.7	22.8
	0.5	✓	✗	36.2	62.8	34.1	23.4
	0.5	✓*	✗	35.8	62.7	34.0	23.2
	0.5	✗	✓	36.3	62.7	33.9	23.2
	0.5	✓	✓	38.0	64.2	35.2	24.3
	0.5	✓*	✓	37.8	64.1	35.1	24.0
CLIP-PGS$_{0.3}$	[0.3, 0.5]	✗	✗	35.9	61.7	33.5	23.0
	[0.3, 0.5]	✓	✗	36.8	63.2	34.3	24.0
	[0.3, 0.5]	✓*	✗	36.7	63.0	34.0	23.9
	[0.3, 0.5]	✗	✓	36.7	63.0	34.5	23.8
	[0.3, 0.5]	✓	✓	38.6	64.4	36.0	25.1
	[0.3, 0.5]	✓*	✓	38.5	64.4	35.7	24.9

Table B9. **Ablation analysis of key components**. We present comprehensive ablation experiments of CLIP-PGS's components, covering zero-shot image classification, linear probing, and text/image retrieval tasks. Here, MR stands for masking ratio, ED for edge detection, and OTN for optimal transport normalization. '*' denotes the use of the Canny edge detection method, Sobel is used by default.

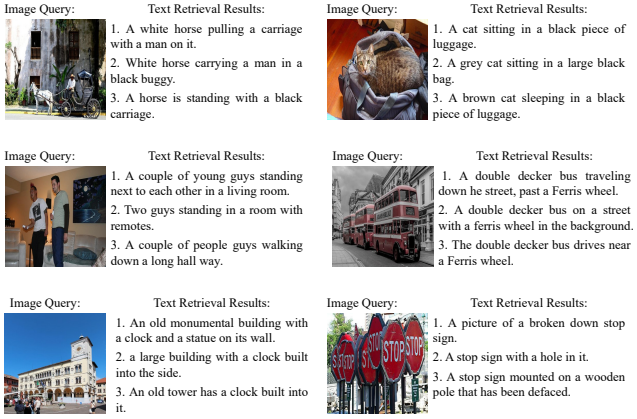

Figure C4. **Visualization of zero-shot image retrieval**. We provide the top-3 predictions of our proposed CLIP-PGS$_{0.3}$. Examples are from the retrieval dataset MS-COCO [18].

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 1

[2] 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*, pages 3558–3568, 2021. 2

[3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017. 1

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 1

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011. 1, 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 2, 3, 4

[8] 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. 1

[9] 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. 1

[10] 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*, pages 178–178, 2004. 1, 2

[11] 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*, 12(7):2217–2226, 2019. 1

[12] 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*, pages 8340–8349, 2021. 1, 2, 4

[13] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1, 2, 4

[14] Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *NeurIPS*, 36:31096–31116, 2023. 2, 4

[15] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013. 1

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 1, 2, 3

[17] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, pages 23390–23400, 2023. 1, 3, 4

[18] 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*, pages 740–755. Springer, 2014. 1, 2, 3, 4

[19] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 1

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 1, 2

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[22] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544, 2022. 1

[23] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian conference on computer vision, graphics & image processing*, pages 722–729, 2008. 1

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 1, 2

[25] 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*, pages 8748–8763, 2021. 1, 3, 4

[26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 1, 2, 4

[27] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *IJCNN*, pages 1453–1460, 2011. 1

[28] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1

[29] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *MICCAI*, pages 210–218, 2018. 1

[30] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 1, 2, 4

[31] Zihao Wei, Zixuan Pan, and Andrew Owens. Efficient vision-language pre-training by cluster masking. In *CVPR*, pages 26815–26825, 2024. 3, 4

[32] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. 1

[33] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. In *ICCV*, pages 2771–2781, 2023. 3, 4

[34] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 1, 2, 3