

1 A Appendix

2 A.1 Additional Examples

3 **Evaluating Fine-Grained Alignment in Image-Text Retrieval.** We compare our trained SuperCLIP-
 4 L 12.8B model with a CLIP-based model (using the current state-of-the-art SigLIP 2 -L as a represen-
 5 tative) to demonstrate the superior performance of our model in fine-grained alignment. Additional
 examples are provided in Fig.1 through Fig.4.

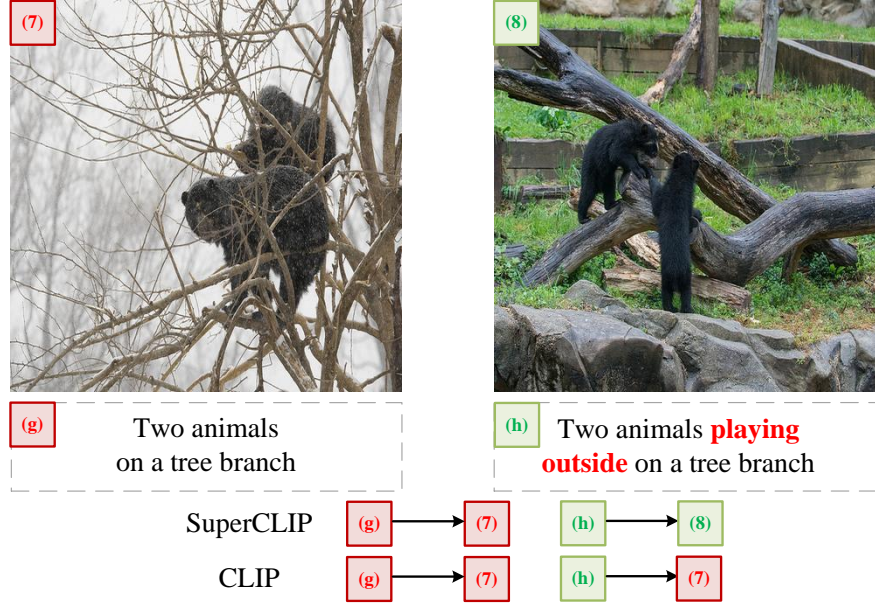


Figure 1: Two animals **playing** outside on a tree branch.

6

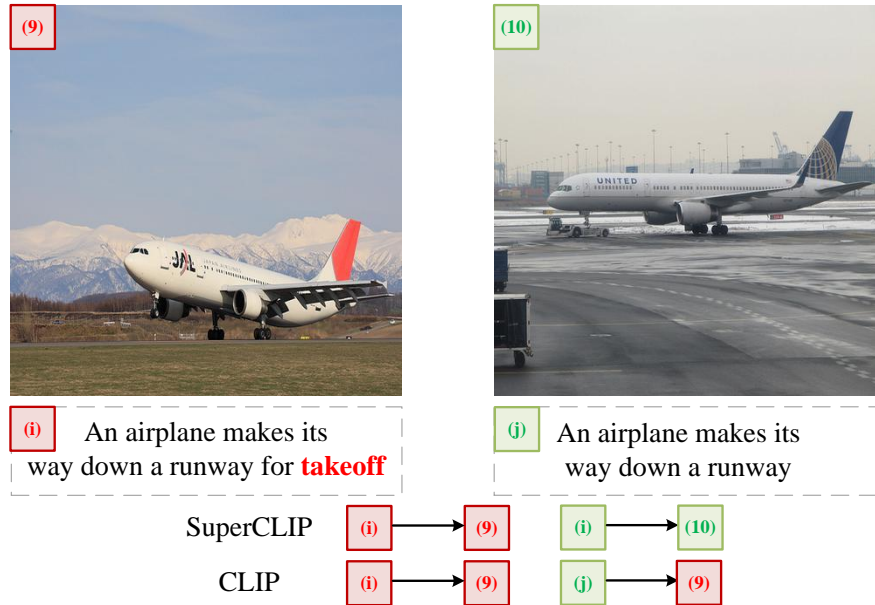


Figure 2: An airplane makes its way down a runway for **takeoff**.

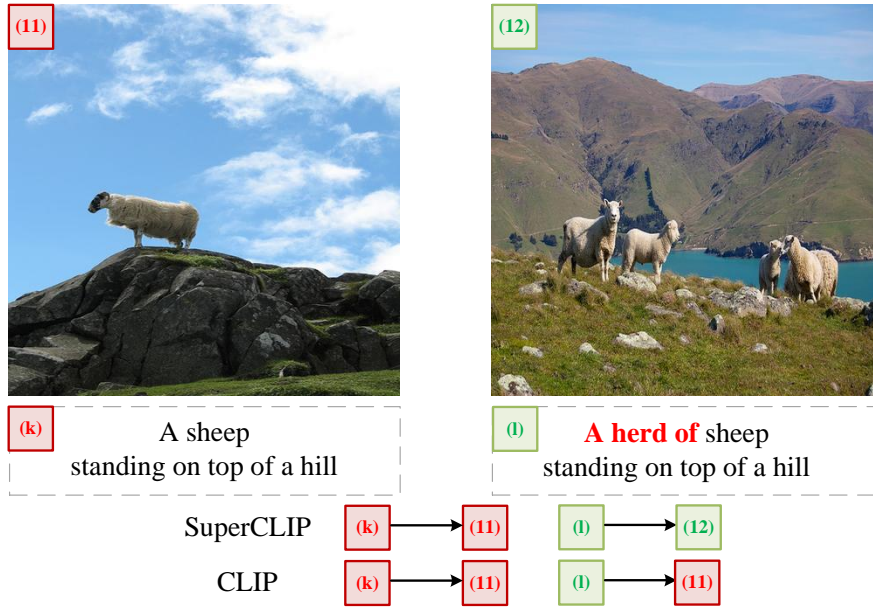


Figure 3: A **herd** of sheep standing on top of a hill.

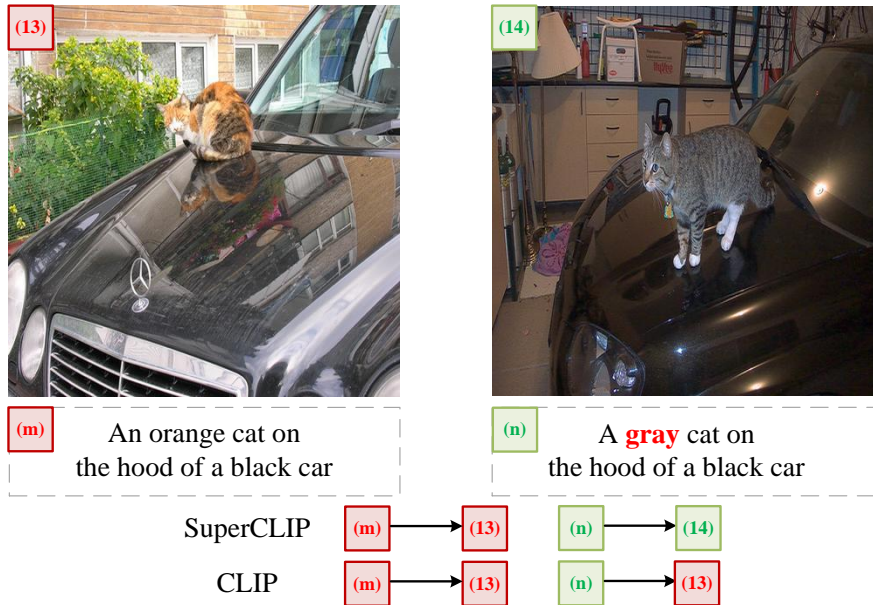


Figure 4: A **grey** cat on the hood of a black car.

7 A.2 Keyword Co-occurrence Statistics

8 We also searched for other keyword co-occurrence statistics in Datacomp-1B (10M captions). Each
 9 example represents a common object pair (**Basic Pair**) combined with a more specific fine-grained
 10 attribute. While the basic pairs tend to appear with relatively high frequency, adding specific fine-
 11 grained attributes results in much lower occurrence rates. More keyword combination examples are
 12 provided in Table 1 through Table 4.

Keyword Group	Animal(s) + Tree		Cat + Car	
	Basic Pair	+ play(ing)	Basic Pair	+ orange/grey
Matching Captions	367	2	216	6
Percentage (%)	0.00367	0.00002	0.00216	0.00006

Table 1: **Keyword combination:** Animals(s) + **Play(ing)** + Tree and **Orange/Grey** + Cat + Car.

Keyword Group	Airplane + Runway		Sheep + Hill	
	Basic Pair	+ takeoff	Basic Pair	+ herd
Matching Captions	58	3	32	1
Percentage (%)	0.00058	0.00003	0.00032	0.00001

Table 2: **Keyword combination:** Airplane + Runway + **takeoff** and **Herd** + Sheep+ Hill.

Keyword Group	Sign + Road		Cat + Toilet	
	Basic Pair	+ yellow/orange	Basic Pair	+ closed/open
Matching Captions	1300	59	68	1
Percentage (%)	0.01300	0.00059	0.00068	0.00001

Table 3: **Keyword combination:** Yellow/Orange + Sign + Road and Cat + **Open/Closed**+ Toilet.

Keyword Group	Man + Tie		Cat + Screen	
	Basic Pair	+ fix(es/ing)	Basic Pair	+ watch/look (ing)
Matching Captions	273	0	38	2
Percentage (%)	0.00273	0.00000	0.00038	0.00002

Table 4: **Keyword combination:** Man + **Fix(es/ing)** Tie and Cat + **Watch/Looking (ing)**+ Screen.

13 As shown in the tables above, common basic pairs such as **Animal + Tree** (Table 1) or **Man + Tie**
 14 (Table 4) appear at a reasonable frequency. However, when more fine-grained attributes are added —
 15 for example, an animal **playing** on a tree or a man **fixing** his tie — such captions become much rarer.

16 A.3 FLOPs Statistics

17 For the FLOPs analysis, the computational cost of the model primarily comes from the **Model**
 18 **Components** and the **Loss Function**. We provide a detailed estimation of the FLOPs for both parts,
 19 as outlined below. The results are summarized in Table 5.

Component	CLIP-B	SuperCLIP-B	CLIP-L	SuperCLIP-L
Vision Encoder	16k * 16.868	16k * 16.868	16k * 59.689	16k * 59.689
Text Encoder	16k * 2.911	16k * 2.911	16k * 6.547	16k * 6.547
Linear Head	–	16k * 0.038	–	16k * 0.051
Contrastive Loss	274.878	274.878	412.317	412.317
Classification Loss	-	5.666	-	5.666

Table 5: **Estimated FLOPs (in GFLOPs)** for different components of CLIP and SuperCLIP under a batch size of 16k. The only additional computation in SuperCLIP comes from the lightweight linear head and the classification loss.

20 **Model Component:** For the CLIP model, the main computational cost comes from the vision
 21 encoder and text encoder. In SuperCLIP, in addition to these two components, a lightweight linear
 22 head is also included, which adds a small amount of computation. We use the standard fvcare library
 23 to compute the FLOPs for the vision encoder, text encoder, and the additional linear head in the CLIP
 24 and SuperCLIP models using dummy inputs. Take the L-size model as an example:

- 25 • The dummy image input is a tensor of shape (1, 3, 224, 224), representing a batch of 1 RGB
 26 image.
- 27 • The dummy text input is a tensor of shape (1, 77), simulating a tokenized text sequence with
 28 77 tokens (the typical maximum sequence length in CLIP), drawn from a vocabulary of size
 29 49,408.
- 30 • The dummy pooled image feature input to the linear head is a tensor of shape (1, 1024),
 31 mimicking the average-pooled image embedding from the vision encoder.

32 In our results, the vision encoder consumes 59.698 GFLOPs, the text encoder consumes 6.547
 33 GFLOPs, and the additional text decoder in SuperCLIP introduces only 0.051 GFLOPs, which
 34 accounts for merely 0.077% of the total model computation.

35 **Loss Function:** To analyze the computational cost of our training objective, we estimate the
 36 number FLOPs involved in the forward pass of the combined loss function $\mathcal{L}_{\text{Total}}$, which includes a
 37 classification loss $\mathcal{L}_{\text{Class}}$ and a contrastive loss $\mathcal{L}_{\text{CLIP}}$. For the classification loss, given a batch size of
 38 B and number of classes C , the total FLOPs include:

- 39 • **log_softmax** over C classes: approximately $5 \times B \times C$ FLOPs, accounting for exponentials,
 40 reductions, and logarithms per sample;
- 41 • **element-wise multiplication** with target distribution: $B \times C$ FLOPs;
- 42 • **summation** across class dimension: $B \times (C - 1)$ FLOPs.

43 Thus, the total cost for classification loss is approximately $7 \times B \times C$ FLOPs. For the contrastive
 44 CLIP loss, assuming image and text features of dimension D , the dominant operation is the pairwise
 45 dot product between B image and B text embeddings, resulting in a matrix multiplication of shape
 46 $[B, D] \times [D, B]$, which requires $2 \times B^2 \times D$ FLOPs. Take the L-size model ($C = 49,408$, $D = 768$)
 47 as an example: with a batch size of 16k ($B = 16,384$), computing the contrastive loss requires
 48 approximately 412.317 GFLOPs, while the classification loss requires only 5.666 GFLOPs—about
 49 1.374% of the former. The difference in computational cost between the two losses mainly arises
 50 from the quadratic complexity of contrastive loss with respect to batch size.

51 A.4 Word-Image Similarity

52 To better illustrate the differences between SuperCLIP and CLIP in terms of word-image similarity,
 53 we present the representative words ranked in the **Top**, **Middle**, and **Last** 20 positions based on
 54 word-image similarity scores. The results are shown in Table 6.

Number	Top 20		Middle 20		Last 20	
	CLIP	SuperCLIP	CLIP	SuperCLIP	CLIP	SuperCLIP
1	zebras	<u>brushing</u>	nintendo	swimming	looks	animals
2	giraffes	<u>monitors</u>	walk	color	skies	snowboarder
3	kites	<u>screen</u>	onto	walking	are	fashioned
4	hydrant	teeth	hot	dining	itself	skiing
5	zebra	wave	roof	sinks	country	skier
6	<u>surfing</u>	<u>sleeping</u>	beans	assorted	leaves	fries
7	<u>skateboarding</u>	racket	alone	motorcycles	types	chain
8	skateboarder	beds	shoes	kite	style	bird
9	surfer	<u>laying</u>	arranged	trail	beneath	jump
10	kite	<u>beside</u>	friends	donuts	about	pasta
11	zoo	desk	women	down	type	zoo
12	surfers	pillow	busy	by	which	railroad
13	<u>brushing</u>	mouth	ledge	eaten	body	puppy
14	snowboarder	<u>inside</u>	plastic	cement	does	snowboarding
15	giraffe	<u>stands</u>	fly	is	professional	skis
16	elephants	<u>blurry</u>	cars	towards	poses	skateboard
17	<u>petting</u>	tarmac	hair	flower	features	skateboarder
18	donuts	sheets	wearing	electronic	watches	tracks
19	carriage	suitcases	helmet	huge	fashioned	giraffe
20	sheep	<u>setting</u>	chair	working	lush	track

Table 6: Representative words from the **Top**, **Middle**, and **Last** 20 positions, selected based on word-image similarity scores computed by CLIP and SuperCLIP on the COCO validation set.

55 As shown in the table above, our model exhibits a significant shift in word ranking compared to CLIP.
 56 While the top 20 words in CLIP are almost exclusively object categories, SuperCLIP successfully
 57 promotes more fine-grained attribute words to higher rank, including those describing object status,
 58 spatial relations, and actions—such as **Blurry**, **Inside**, and **Stand(s)**. This indicates that SuperCLIP
 59 encourages the model to pay more attention to such fine-grained attribute-level concepts.

60 A.5 Complete Evaluation Results Across 38 Datasets

61 We evaluate all models using the open-source LAION CLIP Benchmark on zero-shot classification
62 across 38 datasets. Results are shown in Table 7 (first 19) and Table 8 (remaining 19). As not all
63 datasets are discussed in the main text, we include the remaining references in B.

Model & Mixed	MNIST [18]	VOC2007-Multi [7]	ImageNet-Sketch [29]	SVHN [22]	VOC2007 [7]	Pets [24]	SmallNORB [19]	Rendered SST2 [25]	Diabetic Ret. [11]	DSprites-X [21]	FER2013 [10]	ImageNet-V2 [26]	Caltech-101 [8]	ObjectNet [1]	DSprites-Ori [21]	DTD [4]	SmallNORB-Elev. [19]	PCam [28]	CLEVR-Distance [15]
CLIP-B (1.0/0.0)	44.1	75.1	46.7	39.2	79.6	81.5	5.7	50.0	3.5	3.7	31.4	52.8	84.7	50.0	2.4	41.8	10.9	54.4	18.6
SuperCLIP-B (1.0/0.0)	43.1	78.0	50.9	34.2	80.2	83.4	5.3	52.7	8.4	3.2	33.99	55.4	83.5	54.4	1.5	45.4	11.3	52.1	15.8
CLIP-B (0.0/1.0)	46.3	38.0	19.5	17.0	59.9	19.7	5.7	48.2	28.8	3.1	42.7	18.6	64.3	24.4	1.6	23.4	13.7	56.1	20.7
SuperCLIP-B (0.0/1.0)	60.1	55.2	28.1	15.0	64.5	23.8	5.9	48.6	5.6	3.1	44.3	26.2	68.3	36.2	2.6	26.0	12.5	56.0	17.1
CLIP-B (0.8/0.2)	52.0	75.6	44.9	35.4	79.7	81.1	5.4	48.3	11.1	3.1	39.6	51.2	82.9	49.2	2.5	41.6	11.1	49.9	19.3
SuperCLIP-B (Dual)	64.1	78.5	49.9	37.8	80.7	83.8	5.5	49.8	4.4	3.3	32.2	55.8	83.6	56.9	2.5	43.8	11.1	49.0	18.5
CLIP-L (1.0/0.0)	55.6	77.4	55.2	41.4	80.0	88.2	5.4	53.5	4.8	3.2	33.4	57.6	84.1	56.4	2.2	44.4	12.4	50.5	15.8
SuperCLIP-L (1.0/0.0)	62.3	80.0	59.1	45.7	81.1	90.1	5.3	55.0	6.1	3.1	37.1	62.3	84.9	63.0	1.9	52.4	11.2	50.6	15.4
CLIP-L (0.0/1.0)	37.6	41.7	22.6	26.0	60.0	26.4	5.3	50.3	72.0	3.2	28.4	21.4	66.5	28.5	2.5	24.7	9.8	54.6	15.9
SuperCLIP-L (0.0/1.0)	69.2	56.3	30.3	22.3	67.4	25.6	6.2	52.8	7.0	3.0	44.2	27.8	66.1	44.2	2.5	24.1	9.6	59.6	23.5
CLIP-L (0.8/0.2)	63.3	77.3	50.8	38.8	80.9	83.9	5.5	51.1	59.8	3.2	43.6	56.0	84.2	54.4	2.5	43.1	11.9	50.2	15.6
SuperCLIP (Dual)	74.2	79.7	56.7	47.6	82.3	88.4	5.4	53.7	15.0	2.9	42.4	61.4	84.2	64.8	2.4	48.8	11.2	56.3	19.5

Table 7: Zero-shot classification accuracy (%) on 38 datasets (first 19).

Model & Mixed	ImageNet-R [13]	DMLab [2]	DSprites-Y [21]	EuroSAT [12]	CIFAR100 [17]	FGVC Aircraft [20]	RESISC45 [3]	SUN397 [30]	GTSRB [27]	CIFAR10 [17]	CLEVR-Count [15]	KITTI-Distance [9]	Cars [16]	ImageNet-O [14]	STL10 [5]	Flowers [23]	Country211 [25]	Imagenet-a [13]	Imagenet1k [6]
CLIP-B (1.0/0.0)	66.8	21.3	3.1	51.8	74.8	12.2	48.4	62.3	39.4	93.3	24.9	29.7	75.1	52.2	95.0	60.7	12.6	21.4	60.6
SuperCLIP-B (1.0/0.0)	73.7	19.5	3.2	58.9	76.5	11.8	53.9	64.7	37.8	94.7	20.9	26.2	78.5	47.4	96.5	62.8	13.7	27.8	63.7
CLIP-B (0.0/1.0)	37.2	16.1	3.3	37.3	49.5	1.9	24.6	25.6	13.9	79.7	20.3	39.5	9.1	30.4	83.4	12.5	1.1	6.8	22.4
SuperCLIP-B (0.0/1.0)	50.1	20.9	2.9	46.0	59.4	2.1	38.3	35.1	38.8	86.0	17.4	28.4	15.1	28.1	91.6	13.1	1.5	14.0	30.1
CLIP-B (0.8/0.2)	65.7	18.4	3.2	47.6	73.2	9.7	49.5	62.3	35.6	93.1	25.3	20.4	73.5	52.8	94.8	57.6	11.4	20.5	59.4
SuperCLIP-B (Dual)	74.3	19.9	3.1	48.8	77.3	12.0	56.6	65.4	43.1	95.2	22.5	29.1	76.8	46.2	96.5	62.3	13.5	30.0	63.6
CLIP-L (1.0/0.0)	76.0	18.7	3.2	43.6	79.6	15.9	57.9	66.0	43.2	95.0	17.2	26.0	84.0	47.0	96.9	65.8	15.0	29.4	66.1
SuperCLIP-L (1.0/0.0)	83.1	15.6	3.2	61.3	82.6	16.9	64.2	68.6	49.1	97.2	29.5	22.1	84.6	38.2	98.3	68.9	18.7	41.4	70.2
CLIP-L (0.0/1.0)	42.0	19.7	3.1	44.8	52.8	2.1	35.9	30.5	20.1	82.4	15.2	31.6	10.4	31.9	86.0	11.1	1.5	8.8	25.1
SuperCLIP-L (0.0/1.0)	60.4	20.6	3.1	48.4	61.7	2.4	43.0	40.2	28.5	84.5	27.0	39.0	16.2	24.4	95.9	16.1	1.4	21.5	31.7
CLIP-L (0.8/0.2)	73.1	20.4	3.2	48.4	78.0	11.3	57.5	65.4	43.8	95.5	17.9	25.9	76.2	47.3	96.5	64.8	13.5	28.8	64.0
SuperCLIP (Dual)	82.7	15.8	3.1	58.7	82.3	15.6	62.8	69.1	50.6	97.3	34.7	25.9	81.8	39.8	97.2	66.1	16.5	44.8	68.9

Table 8: Zero-shot classification accuracy (%) on 38 datasets (remaining 19).

64 A.6 Experiments Compute Resources

65 All experiments are performed on a compute cluster with a maximum of 16 NVIDIA A100 GPUs.

66 B References (38 Datasets)

- 67 [1] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund,
68 Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing
69 the limits of object recognition models. *Advances in neural information processing systems*, 32,
70 2019.
- 71 [2] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich
72 Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv*
73 *preprint arXiv:1612.03801*, 2016.
- 74 [3] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification:
75 Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.
- 76 [4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi.
77 Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and*
78 *pattern recognition*, pages 3606–3613, 2014.
- 79 [5] Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in
80 unsupervised feature learning. *Proceedings of the fourteenth international conference on*
81 *artificial intelligence and statistics*, pages 215–223, 2011.
- 82 [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-
83 scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern*
84 *Recognition*, pages 248–255, 2009.
- 85 [7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman.
86 The pascal visual object classes (voc) challenge. *International journal of computer vision*,
87 88:303–338, 2010.
- 88 [8] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few
89 training examples: An incremental bayesian approach tested on 101 object categories. In *2004*
90 *conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.
- 91 [9] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The
92 kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.
- 93 [10] Ian J Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben
94 Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, et al. Challenges in
95 representation learning: A report on three machine learning contests. In *Neural information*
96 *processing: 20th international conference, ICONIP 2013, daegu, korea, november 3-7, 2013.*
97 *Proceedings, Part III 20*, pages 117–124. Springer, 2013.
- 98 [11] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam
99 Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros,
100 et al. Development and validation of a deep learning algorithm for detection of diabetic
101 retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.
- 102 [12] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel
103 dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal*
104 *of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- 105 [13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo,
106 Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness:
107 A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF*
108 *international conference on computer vision*, pages 8340–8349, 2021.

- [14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021.
- [15] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.
- [16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained car images. In *Second Workshop on Fine-Grained Visual Categorization*, 2013.
- [17] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.
- [18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [19] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–104. IEEE, 2004.
- [20] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.
- [21] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. <https://github.com/deepmind/dsprites-dataset/>, 2017.
- [22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.
- [23] 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.
- [24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.
- [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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
- [26] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.
- [27] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.
- [28] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical image computing and computer assisted intervention—mICCAI 2018: 21st international conference, granada, Spain, September 16-20, 2018, proceedings, part II 11*, pages 210–218. Springer, 2018.
- [29] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.
- [30] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.