DIME-FM : Distilling Multimodal and Efficient Foundation Models from CLIP (Supplementary Materials)

Ximeng Sun[1] Pengchuan Zhang[2] Peizhao Zhang[2] Hardik Shah[2] Kate Saenko[1,2] Xide Xia[2]
[1] Boston University, [2] Meta AI

Table 1: **Supplementary Material Overview**

A. Implementation Details

A.1. Training

We pre-compute the teacher features of all images and texts using CLIP-ViT-L/14. We apply the basic data augmentations (only random cropping, flipping and data whitening) to the input images when computing the student features. We empirically find out more advanced data augmentations (such as "rand-m9-n3-mstd0.5") harm the distillation performance. We argue that the advanced data augmentations for the student model's input images enlarge the discrepancy of the student feature and the fixed pre-extracted teacher feature, which lead to the poor performance in vision-language knowledge distillation. We train all models for 100 epochs. During the training, we use the Adam optimizer [16] with decoupled weight decay regularization [22]. We set the initial learning rate as 8×10^{-4} and weight decay 0.05. We warm up the training for 4 epochs and then we decay the learning rate using a cosine schedule [21]. We remove the weight decay of the weights that are the gains or biases. For our model Distill-ViT-B/32, we optimize with the batch size 12288 for images and texts. For Distill-UniCL*, we adopt 8192 as the batch size. We grid search for the best hyperparameters μ^{vl}, $\mu^{p\text{-}vl}$ and μ^{udist} when performing ablation studies on losses in Sec 4.5 (main paper). We finally set $\mu^{vl} = 100$, $\mu^{p\text{-}vl} = 33.3$ and $\mu^{udist} = 14.3$ for all experiments.

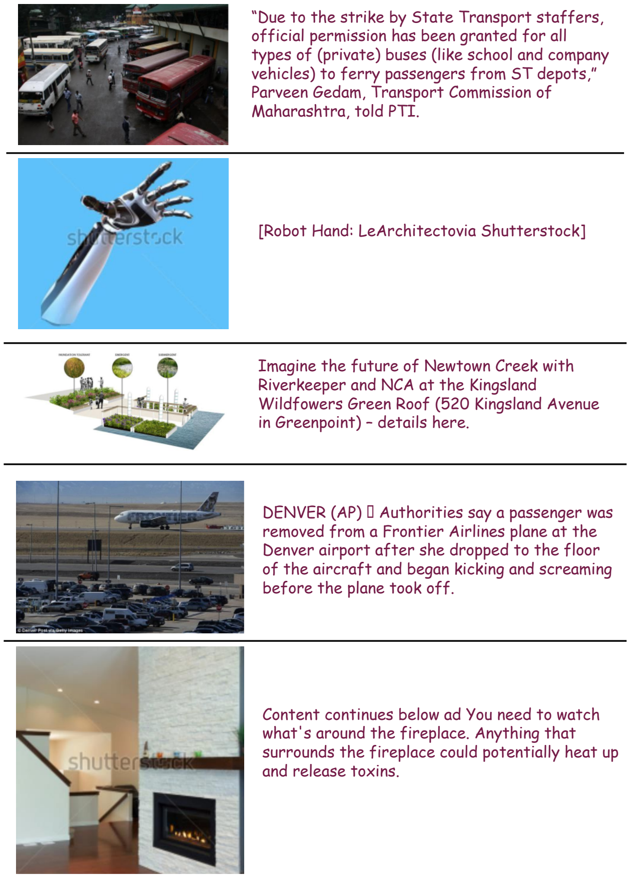
A.2. Evaluation

Transferability to Novel Downstream Tasks. We use ELEVATER toolkit [19] to evaluate the model's zero-shot and linear probing performance on 20 image classification datasets including both coarse and fine-grained tasks: Hateful Memes [15], PatchCamelyon [30], Rendered-SST2 [26], KITTI Distance [9], FER 2013 [14], CIFAR-10 [18], EuroSAT [11], MNIST [6], VOC 2007 Classification [7], Oxford-IIIT Pets [25], GTSRB [29], Resisc-45 [3], Describable Textures [4], CIFAR-100 [18], FGVC Aircraft (variants) [23], Food-101 [2], Caltech-101 [8], Oxford Flowers 102 [24], Stanford Cars [17] and Country-211 [26]. Please refer to Sec.C in Supplementary Material of ELEVATER [19] for detailed dataset statistics and evaluation metrics. We use the same prompt templates as ELEVATER toolkit. We also evaluate the zero-shot performance on ImageNet-1K [5]. For linear-probing performance, we enable the grid search of learning rate and weight decay before finetuning the last classifier layer.

Robustness to Domain Shifts. Following CLIP [26], we use five datasets which have the distribution shifts from ImageNet-1K [5]: ImageNet-V2 Match frequency [27], ImageNet Sketch [31], ImageNet Adversarial [13], ObjectNet [1] and ImageNet Rendition [12]. We use the same 80 prompt templates of ImageNet-1K for these datasets and we report the average zero-shot top-1 performance on these datasets as the metric for Robustness.

B. Visualization of the Constructed $\mathcal{T}$

In Fig. 1, we randomly select several images and visualize their paired sentences of original human annotations and our proposed algorithms. For the second image with robot hand, the chosen sentence describes the image content more accurate than human labels. It indicates our proposed text corpus selection algorithms choose sentences which are not only close to the image in the feature space but also with reasonable concept in the visualization.



"Due to the strike by State Transport staffers, official permission has been granted for all types of (private) buses (like school and company vehicles) to ferry passengers from ST depots," Parveen Gedam, Transport Commission of Maharashtra, told PTI.

[Robot Hand: LeArchitectovia Shutterstock]

Imagine the future of Newtown Creek with Riverkeeper and NCA at the Kingsland Wildfowers Green Roof (520 Kingsland Avenue in Greenpoint) – details here.

DENVER (AP) Authorities say a passenger was removed from a Frontier Airlines plane at the Denver airport after she dropped to the floor of the aircraft and began kicking and screaming before the plane took off.

Content continues below ad You need to watch what's around the fireplace. Anything that surrounds the fireplace could potentially heat up and release toxins.

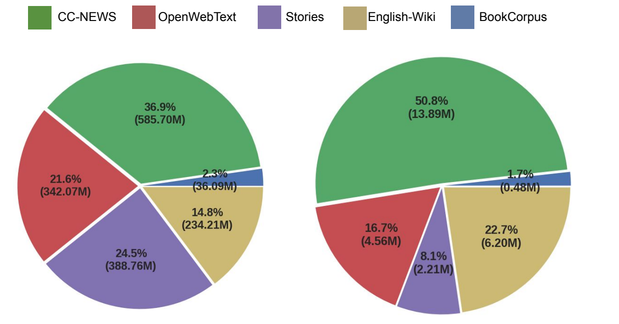
Figure 1: **Selected Sentences from ROBERTa NLP Corpus.**

C. Contribution of each NLP dataset

In Fig. 2 (a) and (b), we counted the contribution of each NLP datasets in original ROBERTa NLP Corpus [20] and our constructed Text Corpus with 28.4M images. Comparing with Fig. 2 (a) and (b), we show our proposed sentence selection algorithm favors sentences from CC-NEWS [10] and English-Wiki, which indicates there are more visually-grounded sentences in these datasets. For instance, they might contain more visual object entities.

D. Paired vs. Unpaired Dataloading

In the main paper, we load in images and texts independently. In this section, we try to load the image and text in pairs (image-caption) with GCC-3M [28], without introducing new losses to use the ground-truth information in the paired image and text data. We re-tune the hyper-parameter for the experiment when loading the paired data. In Table 2, we show loading image and text in pairs does not benefit the transferability during the knowledge distillation and even we observe a small drop when switching to the paired dataloading mechanism. We suspect the drop results from the fewer combinations of image and text seen during the optimization if we load image and text in pairs. It would be an interesting future direction to study how to make better usage of the annotated images and its associating text when there is a small amount of image-text data available, which is out of scope of this paper.



(a) ROBERTa NLP Data Distribution **(b) Constructed Text Corpus Distribution**

Figure 2: **Analysis of Our Constructed Text Corpus from ROBERTa NLP Corpus. Best viewed in color.**

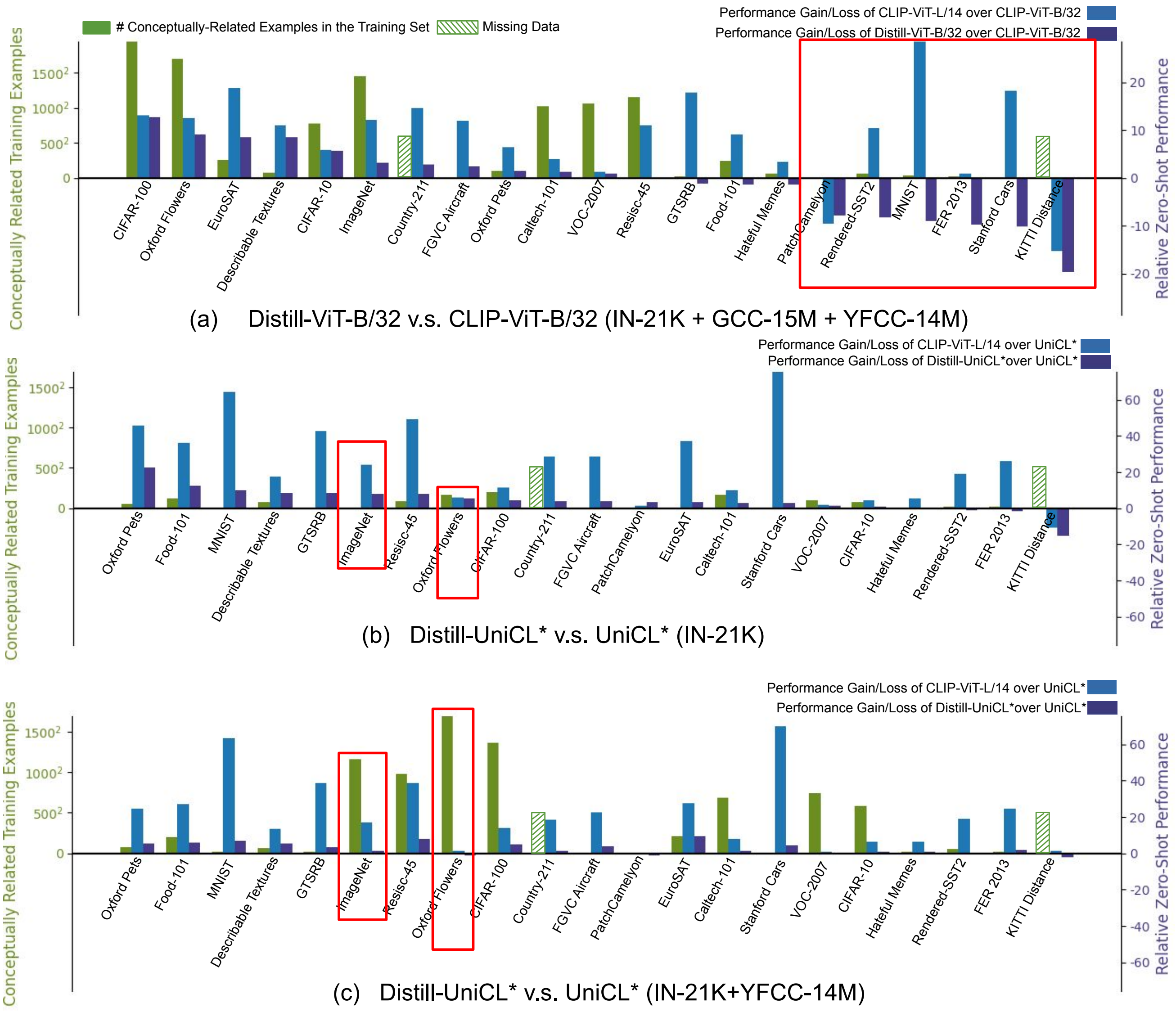
Dataloading	Zero-Shot		Linear Probing
	ELEVATER	IN-1K	ELEVATER
Unpaired	38.6%	39.0%	68.2%
Paired	38.8%	38.3%	68.0%

Table 2: **Paired vs. Unpaired Dataloading.**

E. Conceptual Coverage Analysis

When we evaluate the zero-shot performance of our distilled models on various downstream tasks in ELEVATER and ImageNet-1K benchmarks, we find some tasks benefit more from the teacher model than the others. We study the performance gain or loss of our distilled model for a single downstream task with respect to the teacher model's performance and the number of conceptually-related images available during the training (see Fig. 3). We compute the number of conceptually-related images to each individual downstream task (green bars in Fig. 3) from Table 11 of [32].

- For Fig. 3 (a), we transfer the knowledge from the large teacher model CLIP-ViT-L/14 to our Distill-ViT-B/32 and compare with CLIP-ViT-B/32. Both CLIP-ViT-L/14 and CLIP-ViT-B/32 are trained on the private 400M image-text pairs while Distill-ViT-B/32 is trained on 40M images (consisting of images from IN-21K, GCC-15M and YFCC-14M) and 28.4M unpaired sentences. Our Distill-ViT-B/32 has obvious worse zero-shot performance than CLIP-ViT-B/32 on five out of twenty-one datasets (*i.e.* PatchCamelyon, MNIST, FER-2013, Stanford Cars and KITTI Distance) due to either lack of conceptually-related images in our small training set (MNIST, FER-2013 and Stanford Cars) or the poor performance of the teacher model (PatchCamelyon and KITTI Distance). Notably, if we do not consider PatchCamelyon and KITTI Distance



(a) Distill-ViT-B/32 v.s. CLIP-ViT-B/32 (IN-21K + GCC-15M + YFCC-14M)

(b) Distill-UniCL* v.s. UniCL* (IN-21K)

(c) Distill-UniCL* v.s. UniCL* (IN-21K+YFCC-14M)

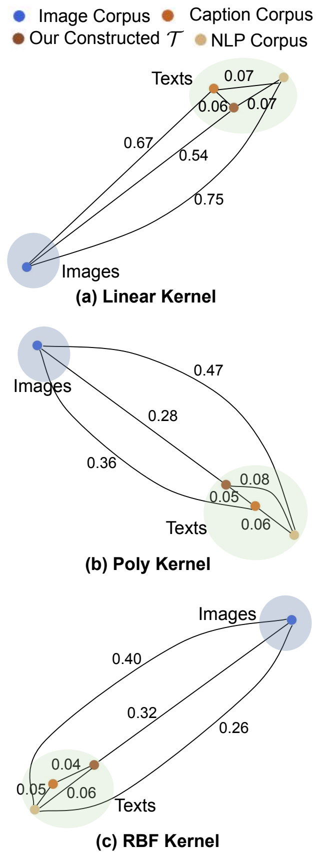
Figure 3: **Conceptual Coverage Analysis of Training Data over Each Downstream Task.**

datasets where the large CLIP model does not full fill a teacher role, our Distill-ViT-B/32 achieves the average Zero-Shot performance 61.03% (vs. 60.95% for CLIP-ViT-B/32) on the remaining 18 datasets in ELEVATER benchmark.

- For Fig. 3 (b) and (c), we transfer the knowledge from CLIP-ViT-L/14 to our Distill-UniCL* and compare with UniCL*. Both Distill-UniCL* and UniCL* are trained on the same relatively-small public datasets besides UniCL* requires the paired image-text data. In Fig. 3 (b), we find distillation from the huge teacher model overall performs better than contrastive pretraining when there are a few conceptually-related training examples. By adding a significant amount (> 1 million) of conceptually-related images (see the comparison of between Fig. 3 (c) and (b) on ImageNet and Oxford-Flowers datasets), the contrastive pretraining gets closer zero-shot performance to our distillation method but the contrastive pretraining requires the additional pairing information from the human annotators. For those datasets which only get fewer than 0.25 million of new images (such as Caltech-101 and EuroSAT datasets), the vision-language distillation still preserves the superior performance to the contrastive pretraining.

F. MMD among Image and Text Corpora

We compute Maximum Mean Discrepancy (MMD) with of four image/text corpora's distributions in CLIP-ViT-L/14's shared feature space using linear, polynomial and RBF kernel respectfully. We use multi-dimensional scaling (MDS) to plot relatively locations of four corpora in a 2D space (see Fig. 4). We have some observations of MMD analysis:



(a) Linear Kernel

(b) Poly Kernel

(c) RBF Kernel

Figure 4: **MMD among four different image/text corpora's in the shared feature space.** We put MMD value on each edge.

- There exists clear modality gap in the shared feature spaces between the image and text features. It consists with the T-SNE plot in Fig. 5 (in the main paper).
- With Algorithm 1, our constructed text corpus is closer to visually-grounded caption corpus than the NLP corpus. It consists with the T-SNE plot in Fig. 4 (in the main paper).
- In the main paper, we show the our selected sentence is closer to the query image that its human-annotated captions. Here, we further show our constructed text corpus is also closer to the image corpus in the distribution level.

G. Full Ablation Studies on Losses

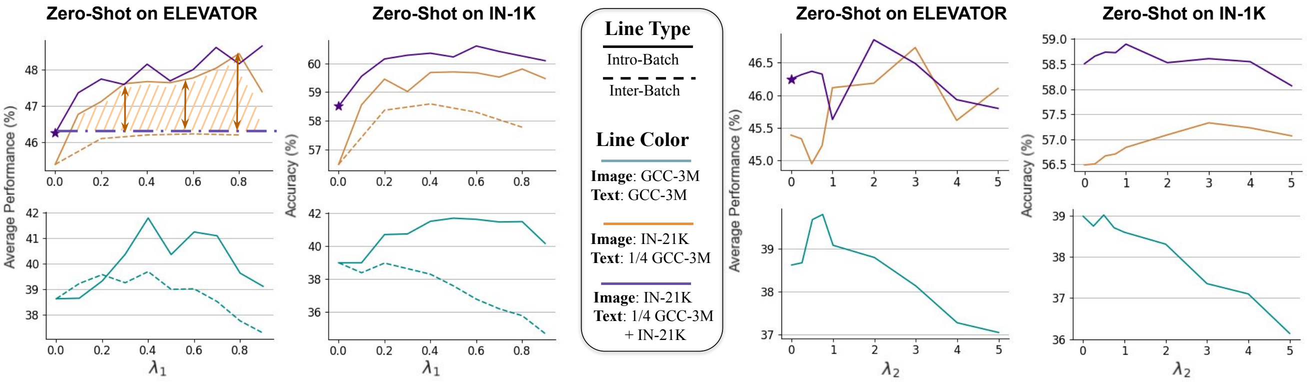
In addition to Sec. 4.5 in the main paper, we provide our full ablation study on losses with different image and text training data. We show the effectiveness of $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$ by Zero-Shot on ELEVATER and IN-1K in three realistic settings (see Fig. 5): (1). The image concepts are entirely overlapped with visually-ground sentences ('Cyan' curves). (2). The image concepts and the visually-grounded sentences are independent ('Orange'). (3). The image concepts are partially covered by the visually-grounded sentences ('Purple').

Ablation on $\mathcal{L}_{p\text{-}vl}$. We gradually put more weights on the $\mathcal{L}_{p\text{-}vl}$ by increasing λ_1 in Eq.6 (main paper) from 0 to 1 in Fig. 5 (a). We compare the inter-batch version (i.e. $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ in Eq.10 from different batches) and intro-batch version (i.e. $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ from the same batch) of $\mathcal{L}_{p\text{-}vl}$ and find the intro-batch $\mathcal{L}_{p\text{-}vl}$ performs better than inter-batch $\mathcal{L}_{p\text{-}vl}$ in setting (1) and (2), so we keep intro-batch version in other experiments. Furthermore, adding $\mathcal{L}_{p\text{-}vl}$ with $\lambda_1 \leq 0.9$ brings better zero-shot performance than only using $\mathcal{L}_{vl}$ in all three settings. However, we observe the dramatic performance drop when we totally replace $\mathcal{L}_{vl}$ with $\mathcal{L}_{p\text{-}vl}$ (*i.e.* $\lambda_1 = 1$). We argue improvement with smaller λ_1's and drop at $\lambda_1 = 1$ both due to the gap between images and text embeddings in the shared feature space. Interestingly, we find that the performance with $\lambda_1 > 0$ in setting (2) is better than using pure $\mathcal{L}_{vl}$ ($\lambda_1 = 0$) in setting (3). It indicates that $\mathcal{L}_{p\text{-}vl}$ is more effective than prompt sentences of class names since using image embeddings as pseudo text embeddings in $\mathcal{L}_{p\text{-}vl}$ introduces richer concepts than class names.

Ablation on $\mathcal{L}_{udist}$. We increase λ_2 from 0 to 5, to introduce $\mathcal{L}_{udist}$ as a regularization term. Generally, $\mathcal{L}_{udist}$ benefits the Zero-Shot on ELEVATER since it tries to preserve the geometry of image features. But different λ_2 works the best for different datasets. $\mathcal{L}_{udist}$ slightly improves IN-1K performance when λ_2 is small but it quickly harms IN-1K performance when λ_2 gets larger. We suspect the poor student embedding in early training along with the large regularization term detours the gradient decent trajectory. $\mathcal{L}_{udist}$ is more effective when the text does not cover the image concepts in setting (2), where it still improves Zero-Shot on ELEVATER and IN-1K with a large λ_2. Our main experiment (Table. 1 in the main paper) further shows $\mathcal{L}_{udist}$ is less effective when applying $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$ together.

H. Detailed Performance on Each Dataset

In this section, we provide the zero-shot performance of models proposed in this paper for each downstream dataset. Other public models' (such as CLIP [26] or UniCL [32]) performance can computed by ELEVATER toolkit [19] (see Table. 3). We also report the number of conceptually-related images computed from Table 11 of [32] for each downstream dataset. Furthermore, we report the detailed performance on five datasets with the domain shift to ImageNet-1K (see Table 4) as the supplement to Table 1 in the main paper.



(a) Ablation Study on Pseudo-MM Score Distillation Loss

(b) Ablation Study on Uni-Modal Distance Preserving Regularizer

Figure 5: **Ablation Studies on Pseudo-VL Score Distillation Loss $\mathcal{L}_{p\text{-}vl}$ and Uni-Modal Distance Preserving Regularizer $\mathcal{L}_{udist}$.** When using IN-21K images as image corpus in both ablation studies, using $\mathcal{L}_{p\text{-}vl}$ and $\mathcal{L}_{udist}$ better utilizes image embeddings. Sometimes, it even performs better than directly incorporating class names in $\mathcal{L}_{vl}$ (i.e the Orange Curve is sometimes above the Purple "– · – ·" line.)

Downstream Tasks	ViT-B/32		Swin-Tiny (IN-21K)			Swin-Tiny (IN-21K+YFCC-14M)		
	# CR-Images	`DIME-FM`	# CR-Images	UniCL*	`DIME-FM`	# CR-Images	UniCL*	`DIME-FM`
Hateful Memes	3.1K	54.4%	80	53.9%	53.1%	322	52.9%	53.6%
PatchCamelyon	158	52.9%	0	49.6%	52.9%	15	51.5%	50.4%
Rendered-SST2	3.4 K	50.0%	650	49.9%	9.4%	3.2K	49.8%	49.8%
KITTI Distance	-	9.28%	-	24.6%	22.2%	-	12.4%	10.3%
FER 2013	579	39.5%	432	24.0%	92.2%	467	25.3%	27.4%
CIFAR-10	0.6M	95.5%	5.9K	91.2%	92.2%	335.8K	89.3%	90.2%
EuroSAT	68.0K	54.0%	0	27.2%	30.4%	46.4K	36.5%	46.1%
MNIST	1.1K	38.8%	0	11.64%	21.6%	619	13.0%	20.1%
VOC 2007 Classification	1.1M	83.4%	3.3K	82.0%	83.3%	544.4K	82.9%	83.4%
Oxford-IIIT Pets	9.1K	88.6%	3.3K	47.6%	70.1%	5.6K	69.1%	74.3%
GTSRB	610	27.6%	0	7.7%	16.2%	545	11.8%	15.0%
Resisc-45	1.3M	55.2%	7.8K	21.6%	29.3%	955.2K	32.3%	40.3%
Describable Textures	5.2K	52.6%	5.2K	37.7%	46.38%	4.4K	42.0%	47.4%
CIFAR-100	3.8M	76.4%	42.2K	66.8%	71.0%	1.9M	64.1%	68.9%
FGVC Aircraft (variants)	90	20.1%	0	3.0%	6.8%	0	9.1%	12.8%
Food-101	57.7K	80.4%	13.8K	57.0%	69.3%	41.9K	65.9%	71.7%
Caltech-101	1.0M	89.6%	28.6K	82.5%	85.2%	475.4%	84.4%	85.7%
Oxford Flowers 102	2.9M	75.9%	26.7K	73.4%	79.0%	2.9M	77.9%	77.0%
Stanford Cars	0	43.35%	0	2.3%	5.0%	0	8.0%	12.5%
Country-211	-	17.1%	-	3.4%	7.5%	-	13.5%	14.9%
Imagenet	2.1M	60.8%	0	51.4%	59.5%	1.3M	58.7%	60.0%

Table 3: **Conceptually-Related Images and Different Models' Zero-Shot Performance on Each Dataset**. We provide the zero-shot performance of every model proposed in this paper for each dataset. We also report the number of conceptually-related images (CR-Images).

Models	ImageNet-v2	ImageNet-R	ObjectNet	ImageNet-Sketch	ImageNet-A	Average
CLIP-ViT-B/32	55.9%	69.0%	**44.2%**	42.3%	31.5%	48.6%
Distill-ViT-B/32 (Captions)	57.5%	69.7%	42.5%	45.7%	31.6%	49.4%
Distill-ViT-B/32 (NLP Texts)	**58.9%**	**69.8%**	43.2%	**46.5%**	**32.2%**	**50.2%**

Table 4: **Robustness.** Our Distilled-ViT-B/32 models perform better than CLIP-ViT-B/32 model on 5 datasets which have distribution shift to origin ImageNet-1K data. The results demonstrates our distilled models preserve the robustness to the distribution shift.

References

[1] 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. *Advances in neural information processing systems*, 32, 2019. 1

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 1

[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] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012. 1

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–308, 2009. 1

[8] 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 *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 1

[9] Jannik Fritsch, Tobias Kuehnl, and Andreas Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 1693–1700. IEEE, 2013. 1

[10] Felix Hamborg, Norman Meuschke, Corinna Breitinger, and Bela Gipp. news-please: A generic news crawler and extractor. In *Proceedings of the 15th International Symposium of Information Science*, pages 218–223, March 2017. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1

[13] 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. 1

[14] Kaggle. Kaggle challenges in representation learning facial expression recognition. 1

[15] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624, 2020. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[19] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, et al. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *arXiv preprint arXiv:2204.08790*, 2022. 1, 4

[20] 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. 2

[21] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 1

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

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

[24] 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. 1

[25] 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. 1

[26] 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. 1, 4

[27] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 1

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 2556–2565, 2018. 2

[29] 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. 1

[30] 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. 1

[31] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[32] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022. 2, 4