Cross-task knowledge distillation for few-shot detection

L. Hémadou[1,2,3] H.Vorobieva [1] A.N. Benaichouche[1] F. Jurie[2] E. Kijak[3]

[1] Safran Tech, Digital Sciences & Technologies Department
[2] Université de Caen Normandie, ENSICAEN, CNRS
[3] Université de Rennes 1, INRIA, CNRS
email: louis.hemadou@inria.fr

Abstract

While powerful pretrained visual encoders have advanced many vision tasks, their knowledge is not fully leveraged by object detectors, especially in few-shot settings. A key challenge in transferring this knowledge via cross-task distillation is the semantic mismatch between outputs: classifiers produce clean probability distributions, while detector scores implicitly encode both class and objectness. To address this, we propose a lightweight fine-tuning strategy guided by a novel, correlation-based distillation loss. This loss aligns the detector's relative class preferences with those of a strong image classifier, effectively decoupling the learning of class semantics from objectness. Applied to a state-of-the-art detector, our method consistently improves performance in a low-data regime, demonstrating an effective way to bridge the gap between powerful classifiers and object detectors.

1. Introduction

Deep learning models typically require large amounts of labeled data to generalize effectively to unseen examples. However, in many practical scenarios, assembling such a dataset can be labor-intensive or impractical. To address this, two main strategies are commonly considered: (1) generating synthetic training data that approximates the distribution of test images, or (2) training directly on a limited dataset. Both approaches, however, present challenges for generalization. Synthetic data may introduce a domain shift with respect to real test images, while training with limited data can hinder the model's ability to learn robust and transferable representations.

In recent years, the emergence of massively pretrained visual encoders [8, 15, 16] has significantly advanced computer vision. Trained on vast corpora, these models excel in domain generalization [21] and few-shot learning [19, 24], making them ideal for data-scarce scenarios. While object detection has benefited from these encoders in open-vocabulary settings [22], their application to standard, closed-vocabulary detection has been less explored. We argue this is due to a fundamental challenge: a direct transfer of knowledge from a pure classifier to a detector is non-trivial due to their incompatible output semantics.

Object detection pipelines combine three subtasks: localization, objectness prediction (does a box contain an object?), and classification (what class is the object?). This decomposition reveals the core of the problem: a classifier's output is a clean probability distribution over classes, whereas a detector's classification scores are implicitly weighted by its objectness confidence. A standard distillation loss like Kullback-Leibler divergence would be confounded by this, penalizing low-confidence predictions even if their relative class scores are correct.

This work directly addresses this semantic mismatch. We propose a lightweight, cross-task distillation framework to enhance few-shot object detectors by leveraging strong pretrained classifiers. Specifically, our contributions are:

- A novel correlation-based distillation loss specifically designed to align the detector's class predictions with a teacher classifier's by focusing on relative score proportionality, thereby ignoring the confounding objectness signal.
- A lightweight fine-tuning strategy to integrate this distillation into any pre-existing detector without requiring architectural changes or expensive retraining from scratch.
- An empirical validation showing that our method consistently improves a state-of-the-art detector (DINO) in a low-data regime, outperforming strong baselines.

2. Related work

Closed-Vocabulary Object Detection. Object detection has evolved through a wide range of architectures, from two-stage methods like R-CNN and its variants [4, 17]

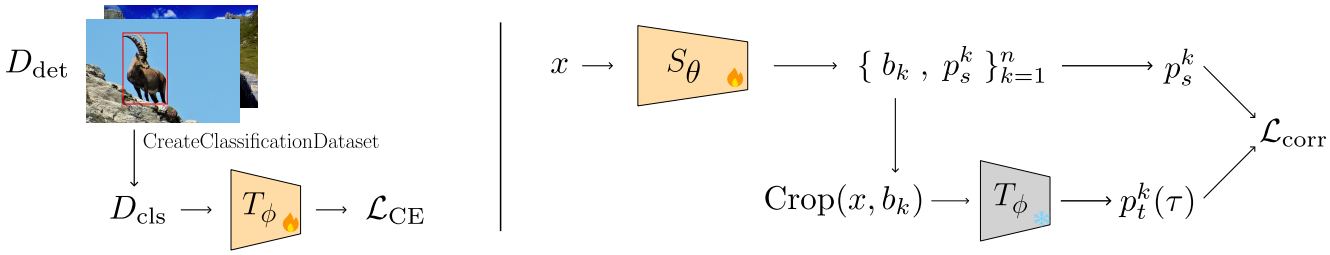

Figure 1. We first extract the crops from the detection dataset D_{det} to create the associated classification dataset D_{cls}, which we use to train a teacher classifier T_ϕ using a standard cross entropy loss. Then, we distil its knowledge into a student detector S_θ by encouraging the detector's class scores p_s^k to be correlated with those of the teacher $p_t^k(\tau)$. To this end, we introduce a novel distillation loss $\mathcal{L}_{\text{corr}}$.

to single-stage approaches [20]. More recently, Detection Transformers (DETR) [1] have established a new paradigm by eliminating hand-crafted components like non-maximum suppression in favor of a set-based prediction framework with Hungarian matching. This has spurred numerous improvements: Deformable DETR [25] improves efficiency and small object handling, while DN-DETR [9] and DINO [23] use denoising objectives to stabilize training and boost performance, particularly in low-data regimes. Given its state-of-the-art performance and stability, we adopt **DINO** as the base detector for our experiments.

Open-Vocabulary Object Detection (OVD). A distinct line of research focuses on OVD, aiming to detect object classes specified by arbitrary text prompts at inference time. These methods typically leverage large-scale vision-language models (VLMs) like CLIP [16]. Prominent strategies include training detectors from scratch on web-scale paired data (e.g., GLIP [10]) or fine-tuning a VLM directly for detection (e.g., OWL [13], OWL-ST [14]). While powerful for generalization, these models can lack the precision of specialized, closed-vocabulary detectors and their fine-tuning on small, specific datasets remains an open question.

Knowledge Distillation for Object Detection. Our work is most related to methods that use knowledge distillation to improve object detectors. Several OVD methods, such as ViLD [5] and DetPro [2], distill knowledge from a frozen VLM into a detector. Their goal is to align the detector's region features with the VLM's text or image embeddings to impart open-vocabulary capabilities. Our approach differs in both goal and mechanism. We focus on the *closed-vocabulary, few-shot* setting, where the aim is to maximize performance on a fixed set of classes with limited data. Instead of distilling from a general-purpose VLM, we perform *cross-task distillation* from a strong image classifier fine-tuned on the target classes. In this respect, our method bears some resemblance to [6], which distills classifier knowledge into a detector via logits- and feature-based losses. However, unlike their approach, we explicitly tackle the semantic gap between classifier outputs and detector predictions by designing a new distillation loss tailored to this cross-task setting.

3. Method

We propose a lightweight fine-tuning procedure to enhance any pretrained object detector by distilling knowledge from a strong image classifier. Our approach introduces a novel cross-task distillation loss designed to handle the semantic differences between the two tasks. The overall process, illustrated in Figure 1, can be decomposed into two parts: (i) training an expert teacher classifier on crops derived from detection annotations, and (ii) fine-tuning the object detector with a small number of additional epochs using our distillation loss to transfer the classifier's knowledge.

3.1. Teacher Classifier Preparation

The first step is to create a high-quality teacher model specialized for the visual domain and class vocabulary of the target detection task.

Let the detection dataset be $\mathcal{D}_{\text{det}} = \{(x^{(j)}, B^{(j)})\}_{j=1}^N$, where $x^{(j)}$ is an image and $B^{(j)} = \{(b_k, y_k)\}$ is a set of ground-truth bounding boxes and their corresponding class labels. From this, we construct a derived image classification dataset, $\mathcal{D}_{\text{cls}} = \{(x_{\text{crop}}^{(k)}, y^{(k)})\}_{k=1}^M$, by extracting the image region $x_{\text{crop}}^{(k)}$ for every ground-truth box b_k across all images.

The teacher classifier, T_ϕ, is constructed by appending a linear classification head, h_ϕ, to a powerful pretrained visual encoder, E_ϕ (e.g., CLIP [16] or DINOv2 [15]). The full set of parameters $\phi = \{\phi_E, \phi_h\}$ is then fine-tuned on $\mathcal{D}_{\text{cls}}$ using a standard cross-entropy loss. Fine-tuning the entire network, rather than just the head, allows the encoder to slightly adapt its general-purpose features to the specific nuances of the objects in our target dataset, resulting in a more expert teacher.

3.2. Cross-Task Distillation with a Correlation Loss

The second stage involves fine-tuning the student detector, S_θ, using a distillation loss that aligns its predictions with the teacher's. For a given prediction from the detector, we

have a class score vector $p_s \in \mathbb{R}^K$ and a bounding box $b \in \mathbb{R}^4$. We crop the corresponding image region and feed it to the teacher classifier to obtain a softened probability distribution $p_t(\tau) = \text{softmax}(T_\phi(x_{\text{crop}})/\tau)$, where τ is the distillation temperature.

Addressing the Semantic Mismatch. A critical challenge arises from the semantic mismatch between the teacher's and student's outputs. The teacher, T_ϕ, produces a clean probability distribution $p_t(\tau)$ (i.e., $|p_t(\tau)| = 1$). In contrast, the student detector's scores p_s often implicitly encode objectness; a prediction with low confidence will have low scores across all classes, and they will not sum to 1. For instance, a classifier might output $[0.9, 0.05, 0.05]$ for classes 'cat', 'dog', 'person'. A detector, however, might predict $[0.8, 0.01, 0.01]$ for a high-confidence cat detection but $[0.1, 0.01, 0.01]$ for a low-confidence one. Consequently, a standard Kullback-Leibler divergence loss would be confounded by this, heavily penalizing the detector for low objectness rather than focusing on the correctness of the class relationships.

The Correlation Loss. To overcome this, we propose a novel loss based on Pearson's correlation coefficient, ρ. Pearson's ρ measures the linear relationship between two sets of data and is invariant to linear transformations (i.e., scaling and shifting). This property is ideal for our purpose, as it makes the loss sensitive only to the relative shape of the score distributions, not their absolute magnitudes. Our loss encourages the student's predictions to be *proportional* to the teacher's:

$$\mathcal{L}_{\text{corr}}(p_s, p_t(\tau)) = 1 - \rho(p_s, p_t(\tau)). \tag{1}$$

This effectively decouples the learning of class semantics from the detector's objectness signal. For detectors that output an additional "no object" or "background" class (e.g., DETR), we simply apply the loss to the K object class scores, ignoring the "no object" logit.

3.3. Two-Stage Training Strategy

Our training process is designed to be efficient and modular.

1. **Stage 1: Detector Pre-training.** The student detector S_θ is first trained to convergence on the target dataset $\mathcal{D}_{\text{det}}$ using only its standard detection loss, $\mathcal{L}_{\text{det}}$. This allows the model to learn the fundamental tasks of localization and coarse classification without the influence of the teacher.
2. **Stage 2: Distillation Fine-Tuning.** We then fine-tune the trained detector for a few additional epochs. In this stage, the model is optimized on a combined objective that includes both the original detection loss and our proposed distillation loss:

Algorithm 1 Cross-Task Distillation Fine-Tuning

Input: Detection dataset $\mathcal{D}_{\text{det}}$, teacher classifier T_ϕ, pre-trained student detector S_θ.
Hyperparameters: Distillation weight λ, temperature τ, sample size m.
for each fine-tuning epoch **do**
  for each image x in a batch from $\mathcal{D}_{\text{det}}$ **do**
    $\{b_i, p_s^i\}_{i=1}^n \leftarrow S_\theta(x)$ *// Get student predictions*
    Compute standard detection loss $\mathcal{L}_{\text{det}}$.
    Let $\mathcal{I} \leftarrow$ Randomly sample m indices from $\{1, \dots, n\}$.
    $\mathcal{L}_{\text{distill}} \leftarrow 0$.
    for $k \in \mathcal{I}$ **do**
      $x_{\text{crop}}^k \leftarrow \text{Crop}(x, b_k)$
      $p_t^k \leftarrow \text{softmax}(T_\phi(x_{\text{crop}}^k)/\tau)$ *// Get teacher prediction*
      $\mathcal{L}_{\text{distill}} \leftarrow \mathcal{L}_{\text{distill}} + (1 - \rho(p_s^k, p_t^k))$
    end for
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{det}} + \lambda(\mathcal{L}_{\text{distill}}/m)$
    Update parameters θ using $\mathcal{L}_{\text{total}}$.
  end for
end for
return Trained detector parameters θ.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{distill}}, \tag{2}$$

where λ is a hyperparameter balancing the two losses. The distillation loss $\mathcal{L}_{\text{distill}}$ is computed as the average $\mathcal{L}_{\text{corr}}$ over a small, randomly sampled subset of m predictions per image to maintain computational efficiency. This two-stage approach enables our method to be a lightweight add-on to any off-the-shelf detector without requiring expensive retraining from scratch. We summarize this stage in Algorithm 1.

4. Experiments

4.1. Experimental Setup

Dataset and Few-Shot Protocol. We conduct experiments on the COCO 2017 dataset [11]. To simulate a few-shot setting, we use a small fraction of the training data, retaining only 2% of the `train2017` split (2365 images). All models are evaluated on the full `val2017` split. For our teacher models, we create a corresponding classification dataset by cropping the object bounding boxes from this 2% split and using their class labels.

Models and Baselines. Our student detector is **DINO** [23] with a ResNet-50 backbone [7], a state-of-the-art DETR-variant. For comparison, we use two primary baselines:

1. **DINO (vanilla)**: The DINO detector trained only on the 2% COCO subset. This is our main baseline to directly measure the benefit of our distillation method.
2. **OWL-ST-FT** [14]: A strong open-vocabulary detector (B/16 and L/14 variants) that we fine-tune on the same 2% data. This baseline assesses how our specialized approach compares to a powerful generalist model in a low-data regime.

Our teacher models are strong image classifiers built by fine-tuning massively pretrained visual encoders (various CLIP [3, 18] and DINOv2 [15] models) on the derived classification dataset.

Training and Distillation Details. We follow a two-stage training process. First, the DINO baseline is trained for 150 epochs on the 2% COCO subset with a learning rate of 1×10^{-4}, followed by a standard learning rate decay. Second, we fine-tune this model for an additional 20 epochs with our proposed distillation loss, using a reduced learning rate of 1×10^{-5}. For distillation, we set the loss weight $\lambda = 1$, teacher temperature $\tau = 3$, and use $m = 20$ predictions per image, selected via a simple grid search. The teacher classifiers are trained to convergence on the classification dataset using the AdamW optimizer [12].

4.2. Main Results

Our Method Boosts Detection Performance. Table 1 presents our main findings. The vanilla DINO baseline, trained on only 2% of COCO, achieves a respectable 21.29% mAP. Applying our cross-task distillation provides a significant and consistent boost. Distilling from a CLIP ViT-H/14 teacher improves performance to 21.98% mAP **(+0.69)**, while a stronger DINOv2 ViT-g/14 teacher pushes the performance to 22.37% mAP **(+1.08)**. This demonstrates that our lightweight fine-tuning strategy effectively transfers knowledge from powerful classifiers to improve the detector's class prediction capabilities.

Comparison with Open-Vocabulary Baseline. Our distilled models also outperform the fine-tuned open-vocabulary baseline, OWL-ST-FT. The best-performing OWL-ST-FT variant (L/14) reaches 20.58% mAP, which is surpassed by both our vanilla DINO baseline and, more significantly, by our distilled models. This suggests that for few-shot, closed-vocabulary tasks, our approach of specializing a strong detector with targeted distillation is more effective than fine-tuning a general-purpose OVD model. We hypothesize this is because DINO's architecture benefits from numerous advances (e.g., query denoising) that are absent in the simpler OWL-ST model.

Better Teachers Lead to Better Detectors. To validate our core hypothesis, we analyze the relationship between teacher quality and student performance. Figure 2 plots the final detection mAP as a function of the teacher's accuracy on the classification task. The results show a clear positive correlation: stronger classifiers consistently lead to larger improvements in the distilled detector. This highlights the importance of the teacher's representational power and confirms that our distillation framework effectively harnesses it.

Table 1. Main results on COCO `val2017` (mAP %) using a 2% training split. Our distillation method significantly improves the DINO baseline and outperforms the OWL-ST-FT competitor. The gain from distillation is shown in red. We did the experiments 5 times (selection of the 2% of COCO, training of the classifier and detector). The observed variance is primarily due to the choice of the training split.

Model	COCO mAP (%)
Open-Vocabulary Baseline	
OWL-ST-FT B/16	18.75 ± 0.28
OWL-ST-FT L/14	20.58 ± 0.26
DINO (vanilla baseline)	21.29 ± 0.33
Our Approach	
DINO + Distill from CLIP ViT-H/14	21.98 ± 0.33 **(+0.69)**
DINO + Distill from DINOv2 ViT-g/14	22.37 ± 0.31 **(+1.08)**

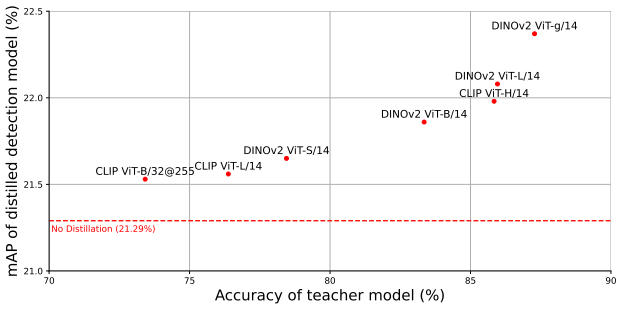

Figure 2. Final detection mAP vs. teacher model accuracy. There is a strong positive correlation, showing that better classifiers lead to better-distilled detectors.

5. Conclusion

We introduced a cross-task distillation strategy to enhance few-shot object detectors. By using a novel correlation-based loss, our method effectively transfers knowledge from a strong classifier by aligning relative class scores, overcoming the semantic mismatch between the two tasks. This lightweight fine-tuning approach yields consistent performance gains on a state-of-the-art detector, demonstrating a practical way to leverage large-scale pretrained encoders for closed-vocabulary detection.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, pages 213–229. Springer, 2020. 2

[2] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14064–14073. IEEE, 2022. 2

[3] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T. Toshev, and Vaishaal Shankar. Data filtering networks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 4

[4] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society, 2014. 1

[5] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. 2

[6] Shuxuan Guo, José M. Álvarez, and Mathieu Salzmann. Distilling image classifiers in object detectors. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1036–1047, 2021. 2

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 3

[8] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment. 2024. 1

[9] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. DN-DETR: accelerate DETR training by introducing query denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13609–13617. IEEE, 2022. 2

[10] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022. 2

[11] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 3

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 4

[13] Matthias Minderer, Alexey A. Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part X*, pages 728–755. Springer, 2022. 2

[14] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 2, 4

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. 1, 2, 4

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 1, 2

[17] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99, 2015. 1

[18] Alex Fang Jonathan Hayase Georgios Smyrnis Thao Nguyen Ryan Marten Mitchell Wortsman Dhruba Ghosh Jieyu Zhang Eyal Orgad Rahim Entezari Giannis Daras Sarah Pratt Vivek Ramanujan Yonatan Bitton Kalyani Marathe Stephen Mussmann Richard Vencu Mehdi Cherti Ranjay Krishna Pang Wei Koh Olga Saukh Alexander Ratner Shuran Song Hannaneh Hajishirzi Ali Farhadi Romain Beaumont Sewoong Oh Alex

Dimakis Jenia Jitsev Yair Carmon Vaishaal Shankar Ludwig Schmidt Samir Yitzhak Gadre, Gabriel Ilharco. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 4

[19] Haoyu Song, Li Dong, Weinan Zhang, Ting Liu, and Furu Wei. CLIP models are few-shot learners: Empirical studies on VQA and visual entailment. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6088–6100. Association for Computational Linguistics, 2022. 1

[20] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9626–9635. IEEE, 2019. 2

[21] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7949–7961. IEEE, 2022. 1

[22] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14393–14402. Computer Vision Foundation / IEEE, 2021. 1

[23] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 2, 3

[24] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 1

[25] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2