Zero-Shot Distillation for Image Encoders: How to Make Effective Use of Synthetic Data

Niclas Popp[1,2], Jan Hendrik Metzen[2], Matthias Hein[1]

[1]University of Tübingen
[2]Bosch Center for Artificial Intelligence (BCAI), Robert Bosch GmbH

Abstract. Multi-modal foundation models such as CLIP have showcased impressive zero-shot capabilities. However, their applicability in resource-constrained environments is limited due to their large number of parameters and high inference time. While existing approaches have scaled down the entire CLIP architecture, we focus on training smaller variants of the image encoder, which suffices for efficient zero-shot classification. The use of synthetic data has shown promise in distilling representations from larger teachers, resulting in strong few-shot and linear probe performance. However, we find that this approach surprisingly fails in true zero-shot settings when using contrastive losses. We identify the exploitation of spurious features as being responsible for poor generalization between synthetic and real data. However, by using the image feature-based $\mathcal{L}_2$ distillation loss, we mitigate these problems and train students that achieve zero-shot performance which on four domain-specific datasets is on-par with a ViT-B/32 teacher model trained on DataCompXL, while featuring up to 92% fewer parameters.

Keywords: Data-Free Knowledge Distillation, CLIP, Synthetic Data, Zero-Shot Classification

1 Introduction

Motivation. Image classifiers built on top of large vision(-language) foundation models, such as CLIP [35] or DINOv2 [32], have shown impressive zero-shot capabilities across various tasks. However, their extensive parameter count and high inference latency present significant challenges for deployment in resource-constrained edge devices used in driver-assistance systems, automated driving, mobile robotics, or video surveillance. Due to their reduced capacity, smaller models cannot be expected to match the performance of larger ones in arbitrary domains. Additionally, training large-scale foundation models typically involves several millions or billions of images, making it expensive and time-consuming. Together, this motivates the need for smaller domain-specific models, as well as data-efficient training procedures. In this work, we specifically focus on zero-shot image classification, for which only a small-scale image encoder is required. Class-specific text embeddings are fixed and can be precomputed off-device, while only image embeddings are computed on-device. Thus, our goal is to distill

smaller drop-in replacements (students) of the CLIP image encoder (teacher) that achieve on-par performance on the specific target domains of interest. In particular, we want to specialize the image encoder student to novel domains for which we only know the relevant classes, but do not have access to actual images, the so called *zero-shot distillation* setting.

For zero-shot distillation, domain-specific data can be obtained from "general-purpose" generative models, such as large-scale latent diffusion models [34], by class-aware prompting. However, learning from synthetic images has proven challenging [2, 37]: using simple class-specific text prompts, like those used for zero-shot classification by CLIP [35], yields low-diversity datasets and poor classifiers, as observed in previous studies [2, 37]. Furthermore, when training on a combination of natural and such low-diversity synthetic images, the overall accuracy starts to decline as synthetic data outweighs real data [2]. More advanced methods for diversifying prompts with large-language models [51], together with a large compute budget for synthetic image generation, allow achieving high accuracy in linear probe or few-shot scenarios, indicating strong representation learning capabilities. However, we observe in this work that even with diverse prompting, the performance in zero-shot distillation with actual zero-shot evaluation remains comparatively low. We identify that fine-tuning small CLIP image encoders with contrastive text-image losses leads to models that exploit spurious features in images and because of this exhibits poor generalization between synthetic and real images. Linear probe evaluation involves a final linear layer trained on real-data and is thus not affected by the same issue.

Besides identifying the issue of spurious feature learning, our main contribution is the (somewhat unexpected) observation that a simple change of the distillation loss can mitigate this problem. Specifically, we find that employing a $\mathcal{L}_2$ loss between student's and teacher's image features substantially reduces the tendency of the models to exploit spurious features and enhances their generalization capabilities between synthetic and real data. We attribute this to the fact that this loss distills image encoders without any influence of the teacher's text encoder and the potential shortcut learning it might encourage. Through one epoch of pre-training on DataComp medium [7] and subsequent fine-tuning on diverse synthetic datasets generated using diffusion models and prompts from large language models, we achieve superior zero-shot classification performance on four target datasets compared to TinyCLIP [48], the current state-of-the-art for distilled CLIP models. Furthermore, our approach achieves on-par performance with the teacher model, all without utilizing a single annotated image from the target domain during training. When using a image encoder that features only 11 million trainable parameters, we manage to achieve the zero-shot performance of a ViT-B/32 with 86 million parameters within a margin of 5 percent points and outperform a TinyCLIP model with 90 million trainable parameters on three of four test datasets.

Main contributions. In this work, we show that small 1-to-1 replacements of the CLIP vision encoder can be efficiently and robustly trained in a zero-shot setting using feature distillation. Our main contributions are the following:

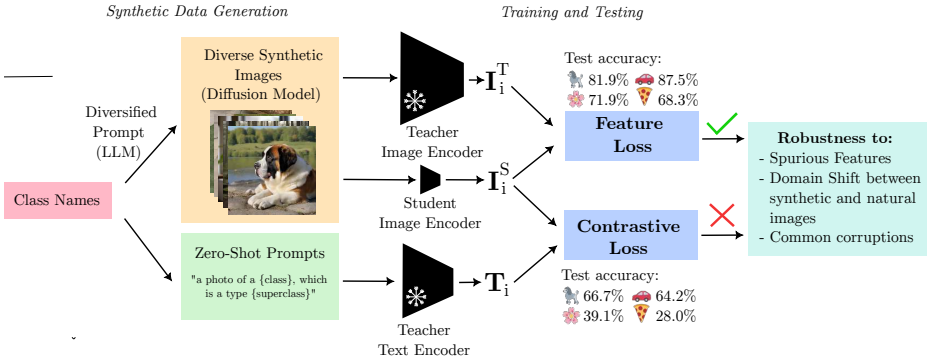

Fig. 1: Overview over our zero-shot distillation framework and the observed properties of the contrastive and feature losses. The reported test accuracies are from a fine-tuned TinyViT-11M vision encoder.

1. We introduce a unifying framework for zero-shot distillation of image encoders.
2. We identify failure cases of distillation when using contrastive losses for fine-tuning on small or synthetic datasets. We attribute this to models being prone to learning spurious features and overfitting to the training domain.
3. We propose feature distillation, which is not susceptible to these spurious features and generalizes better between synthetic and natural images.
4. Using our framework we manage to distill a ViT-B/32 CLIP vision encoder into student models with up to 93% fewer parameters that closely match the classification performance of the teacher and surpass existing baselines on the Oxford Pets [33], Flowers-102 [30], Stanford Cars [16] and Food-101 [3] datasets.

2 Related Work

Knowledge Distillation of Vision-Language Models. Knowledge Distillation [14] is a widely used technique for transferring knowledge from larger teacher models to smaller student models. In its vanilla form, the approach involves combining a standard training loss with a distillation loss that considers the output of both the student and teacher models, penalizing discrepancies between the two models. Knowledge distillation has been observed to not only benefit the test accuracy of the student on the target datasets but transfer other favorable properties of the teacher such as domain generalization [31]. While this approach has been well-established for single-modality tasks including vision [27, 49] or language [15, 36], recent works have extended the concept to the multi-modal setting, specifically in the context of vision-language models. CLIP-KD [50] provides an extensive set of experiments comparing various different loss combinations. TinyCLIP [48] proposed an advanced initialization process using weight inheritance from the teacher to the student as well as a multi-stage progressive

distillation culminating in models that are only 1/4 the size of a ViT-B/32 CLIP model. MobileCLIP [46] further refined the distillation process by incorporating image augmentation, synthetic captions, and dedicated architectural choices. In contrast to these existing methods, our approach focuses on finding only a one-to-one replacement of the vision encoder while the text encoder remains frozen. Apart from CLIP-specific techniques, unsupervised distillation based purely on images without labels has been identified as a data-efficient alternative to supervised training for vision encoders [12] and class-incremental generalization [18]. We build on this observation by combining unsupervised pre-training with targeted fine-tuning. Despite knowledge distillation being a widely adopted training technique, it has been observed that it does not always work as commonly understood. Even when the student features the same capacity as the the teacher, there can be significant discrepancies in their predictive distributions [31, 39].

Training and Distillation Using Synthetic Images. Recent advancements in generative text-to-image models have sparked a growing interest in utilizing synthetic images for vision applications. Azizi et al. [2] demonstrated that images from fine-tuned text-to-image models can be combined with real images to enhance the accuracy of classifiers on ImageNet-1k [5]. For text-to-image generation, diffusion models are commonly employed, particularly for knowledge distillation [19]. However, it was observed that the performance deteriorates when the number of synthetic images surpasses that of real images. Yu et al. [51] attributed this decline to the lack of diversity in the used synthetic images. To mitigate this issue, they proposed a strategy to diversify the image generation process by incorporating prompts generated by large language models, thereby enhancing content and style variation. Another approach to diversification in the few-shot setting was presented by Da Costa et al. [4], which involved augmentations and low-rank adaptation. By scaling up synthetic datasets, Tian et al. [42] and Hammoud et al. [11] demonstrated the feasibility of training vision-language foundation models solely using images from text-to-image models. However, achieving performance on par with or surpassing models trained on real data necessitates the utilization of a large number of synthetic images, on the order of 10^7 or 10^8. This not only prolongs the already long training process but also introduces additional computational overhead. Most importantly, the reported results are typically obtained by linear probing or after few-shot training and are not true zero-shot accuracies which we aim to optimize. To mitigate these challenges, we propose a hybrid approach that combines a large-scale dataset of natural images with a smaller set of domain-specific synthetic images.

3 Framework for Zero-Shot Distillation

Zero-shot distillation refers to the process of transferring knowledge from a teacher to a student model in a setting where one does not have access to images from the target domain. It is thereby a special case of data-free knowledge distillation which describes the setting where the training data for teacher and the student differ. Zero-shot distillation specifically focuses on the ability of founda-

tion models as teachers to perform well on unseen data due to their generalization properties. The objective is to transfer this performance to a smaller student model without utilizing any of the unseen data. Therefore, the primary goal is not to address the disparity between the datasets used to train the teacher and student, but rather to extract domain-specific knowledge from the teacher model without having access to the corresponding data. The term zero-shot distillation has been introduced previously [28], yet only in the setting for single-modal classifiers that were trained using the cross-entropy loss. In our case, we consider CLIP which is a vision-language model instead of a simple image classifier. In this section, we present a structured framework for zero-shot distillation. Specifically, we discuss the data domains, the training pipeline, the generation of diversified synthetic training data as well as the selection of an appropriate loss function.

3.1 Data Domain

In the context of (pre-)training for a zero-shot setting, there are currently two core approaches. The first one involves relying on large-scale data such as common crawl datasets [7, 38]. While this approach is feasible for large foundation models, it poses challenges for smaller models as these lack the same level of generalization capabilities due to their smaller capacity. The second approach involves training from scratch using either purely synthetic images [11, 42, 43] or a combination of real and synthetic images [2, 51]. Yet by incorporating few-shot learning [11, 43] on real images or linear probing [11, 42, 43] after training on synthetic data, the reported accuracies are no longer truly zero-shot. We optimize the actual zero-shot performance by adopting a two-stage approach: in the first stage, we pretrain on a large-scale general-purpose dataset consisting of natural images, and subsequently fine-tune using a smaller set of domain-specific synthetic images. This approach allows us to address the limitations of relying solely on generalization or training on synthetic data, and enables us to achieve strong zero-shot performance.

3.2 Training Pipeline

In order to shorten training in comparison to training from scratch, Wu et al. [48] have introduced weight inheritance as an initialization scheme for distilling CLIP models. This method has a significant limitation as it can only be applied when the student model shares a similar architecture with the teacher model. Instead of using weight inheritance, we introduce a pre-training step [8] which is not targeted to a specific domain. Pre-training as for large foundation models like the original CLIP [35] typically requires substantial computational resources due to the use of billions of images. This contradicts the objective of resource-constrained training for small models. He et al. [12] observed that pre-training can be shortened significantly by using a feature-based loss. For our purpose of training 1-to-1 replacements of CLIP vision encoders, this step has further advantages: by aligning the embeddings of teacher and student, we can mitigate phenomena like the modality gap [20] where corresponding output vectors are

located in different areas of the embedding space. Subsequently, we fine-tune the pre-trained models towards the target domain of interest with the same loss. The only difference between pre-training and fine-tuning is the training data.

3.3 Data Diversification of Synthetic Images

In the context of zero-shot learning for image classification, synthetic data generation is based on the class names. However, it has been observed that using only the names to generate images using diffusion models leads to suboptimal performance [37]. This is primarily due to the lack of diversity in the generated images as well as class ambiguity [4]. One alternative is to utilize captions from existing real datasets for image generation. This approach is not entirely consistent with a zero-shot setting, as there may not be available captions for all classes. To address this challenge, recent approaches have turned to leveraging large language models (LLMs) to enhance diversity in the prompts. In addition to class names, LLMs are guided by additional inputs for diversification, such as information from a concept bank [11] or specific requirements related to contextual and style diversification [51]. By incorporating these additional sources of guidance, the generated synthetic data becomes more diverse and aligned with the desired objectives of the target setting. Using the approach from Yu et al. [51], we focus on *contextual dimensions* to achieve diversification. These dimensions are attributes that describe the context of the image such as the background, camera angle, object position, presentation style, and superclasses, all of which are tuned specifically for the target dataset. In contrast to Yu et al. [51], we do not prompt the LLM for each caption separately, but ask for different options for each contextual dimension. This reduces the risk of obtaining similar captions. The final prompt used for the text-to-image model is a comma-separated list of options for these contextual dimensions. Instead of using all possible combinations of options, which would result in a strongly growing number of images given more options, we perform combinatorial testing [1, 29].

3.4 Loss Selection

Knowledge distillation involves combining a training loss $\mathcal{L}_{training}$ with a distillation loss $\mathcal{L}_{distillation}$ [14]. The training loss $\mathcal{L}_{training}$ takes into account the image and ground truth labels or captions, while the distillation loss $\mathcal{L}_{distillation}$ is used to align the teacher and student models. Commonly, the overall loss is selected as $\mathcal{L}_{overall} = \mathcal{L}_{training} + \lambda \mathcal{L}_{distillation}$, where λ is a scaling parameter [9]. One might assume that using the CLIP loss for $\mathcal{L}_{training}$ would be the most direct approach, as it was used to train the teacher model. However, He et al. observed that pre-training can be substantially sped up by solely employing a feature-based distillation loss without a supervised training loss. In contrast to the CLIP loss, which aims at aligning the text and image embedding of image-caption pairs, the student directly learns from the image features of the teacher without considering the text. More precisely, for every optimization step, we sample a batch of N images $\{\mathbf{t}_i, \mathbf{i}_i\}_{i=1,\dots,N}$ as input. Denote the normalized image

embedding corresponding to the i-th image by $\mathbf{I}_i^S$ for the student and $\mathbf{I}_i^T$ for the teacher. Based on this, the $\mathcal{L}^2$ feature distillation loss $\mathcal{L}_2^{\text{feature}} = \sum_{i=1}^N \|\mathbf{I}_i^S - \mathbf{I}_i^T\|_2$ is optimized. For fine-tuning, the CLIP loss function remains the commonly used approach. [10]. Yet, it was initially designed to pre-train both the text and image encoder on large-scale vision-language datasets where each image has a distinct caption, for most image classification datasets, each image only possesses a class label instead of a diverse prompts. In this case, we employ the zero-shot captions "a photo of {class name} which is a type of {superclass}", which were originally introduced for the zero-shot inference of the original CLIP model [35], as $\mathbf{T}_i$. "Superclass" refers to a general description of the object that can be encountered such as pets, food, cars or similarly. By using these class-specific prompts, several images in a batch may share the same caption which conflicts the goal of decreasing the similarity of images embeddings to the text embeddings of not matching captions in the CLIP loss. An alternative to the CLIP loss is given by the multi-positive contrastive loss introduced in StableRep [43]. To adapt the multi-positive (MP) loss to our setting, we replace the anchor sample by the embedding of a class-specific zero-shot prompt. Denote by $\mathbf{Z}_k$ the normalized embedding of the zero-shot prompt for class k. The contrastive distribution is given by $q_k = \frac{\exp(\langle \mathbf{I}_i, \mathbf{Z}_k \rangle / \tau)}{\sum_{j=1}^M \exp(\langle \mathbf{I}_i, \mathbf{Z}_j \rangle / \tau)}$ and the ground-truth categorical distribution is $p_k = \frac{\mathbb{1}_{l(\mathbf{I}_i)=k}}{\sum_{j=1}^M \mathbb{1}_{l(\mathbf{I}_i)=j}}$. Given M classes, the MP loss is computed as $\mathcal{L}_{\text{MP}} = -\sum_{k=1}^M p_k \log q_k$. We compare contrastive loss, MP loss, feature loss, as well as combinations thereof, on both synthetic and real data in our experiments. We observe that in the zero-shot settings, pure feature distillation is both the most efficient choice for pre-training and the most robust loss for fine-tuning.

4 Experiments

In this section, we present the results of the models trained based on our framework and thereby explain how feature distillation enables zero-shot distillation with synthetic data. After a description of the setup we will compare our models to existing baselines which we outperform consistently. In the ablation studies, we uncover that using contrastive losses leads to students that exploit spurious features in the data, generalize poorly between synthetic and real data alongside being less robust to common corruptions of input images. Overall, this indicates that contrastive losses are a potential cause why efficient zero-shot distillation from synthetic has been an unresolved challenge.

4.1 Setup

Datasets and Hyperparameters As introduced in Section 3.2, we perform feature-based pre-training on a large-scale dataset consisting of natural images for various domains. For this purpose, we select DataComp medium [7] and train

for a single epoch. Originally, the dataset consists of 123 million images but at the time we conducted out experiments only 86% of the image URLs were still active. For fine-tuning, we target the Oxford Pets [33], Oxford Flowers [30], Food-101 [3] and Stanford Cars [16] to evaluate or models domain-specific datasets. These datasets are only used for testing while the actual training datasets used for training are synthetically generated based on the classes. Using the diversification strategy discussed in Section 3.3, we select a different set of five contextual dimensions and corresponding weights in the prompts to the diffusion model. More details on this selection are given in the supplementary material. For the pets, flowers and cars dataset we generate 15 options per contextual dimension while for the food dataset we use 30 as the target dataset is larger as well. This results in 265 and 1011 images per class, respectively. As the selection of options for the contextual dimensions and superclasses are relatively simple, we can use a smaller language model Llama-2 7B fine-tuned for chats [45] and still obtain sufficiently diverse prompts. For the generation of the images, we utilize a LCM LoRA [25] of Stable Diffusion XL [34] with a guidance scale of 0.5 and prompt weighting. As in the original CLIP paper [35] random square crops of the resized images are the only data augmentation used during training.

For both pre-training and fine-tuning we use the same hyperparameters. We train using a batch size of 256 and a constant learning rate of 5×10^{-4} for the AdamW optimizer [24]. All other hyperparameters were kept consistent with the CLIP training methodology [35]. One epoch of pre-training corresponds to 4.3×10^5 optimization steps. For fine-tuning, we perform 96 optimization epochs for all models. Even on the largest synthetic dataset this equals less than 9% of the update steps done during pre-training.

Student and Teacher Architectures As teacher model, we employ a ViT-B/32 [6] CLIP vision encoder that has been trained on DataComp-XL, a dataset consisting of 12.8 billion image-text pairs from Common Crawl [7]. The corresponding text encoder follows the same architecture as described in the original CLIP paper, with 63 million parameters [35] and an embedding dimension of 512. For our student models, we utilize two different types of architectures: EfficientNets [40], which are based on convolutional neural networks, and TinyViTs [40], which are hybrid models combining convolutions and transformers. For our final results, we respectively select three models in the 5, 10, and 20 million parameter range from each architecture type. We only report intermediate results on the TinyViT with 11 million parameters. All models are systematically scaled down to improve inference speed and reduce the number of parameters while still maintaining a high capacity for representation learning. To align the output of the vision encoder with the embedding dimension of the teacher model, we apply a single linear projection head. This ensures consistent dimensionality of the embedding space between the teacher and student models.

4.2 Zero-Shot Performance and Comparison to Baselines

In this section, we report the zero-shot classification accuracy of our models based on the TinyViT-11M architecture and compare them to existing bench-

marks. The results are shown in Table 1. The first reference for the performance is the teacher itself which is trained on DataComp-XL as well as the same model trained on DataComp-medium. The teacher model achieves zero-shot accuracies of over 80% on the pets, cars and food datasets as well as over 70% on the flowers dataset. Additionally, we report the performance of four TinyCLIP models that have been trained on LAION-400M [38] or YFCC-15M [41] datasets. These TinyCLIP models have undergone extensive training on large-scale datasets for multiple epochs, which differs from our approach. Specifically, the smallest TinyCLIP model has been exposed to six times as many images as our models, while the largest models have encountered over 120 times as many samples. The TinyCLIP models exhibit a comparable size to our models in terms of the number of parameters, even when considering the frozen text encoder which is not required for zero-shot classification. The largest TinyCLIP model has 40% fewer parameters than the ViT-B/32 CLIP model and achieves its performance up to a margin of 9%. The smallest TinyCLIP model features the same number of trainable parameters but has a gap of over 75% to ViT-B/32 CLIP model on the cars dataset. At the time of conducting our experiments, we were unable to compare our results with MobileCLIP [46] and CLIP-KD [50] as these models are not publicly available.

We report three types of models from our framework: trained from scratch, pretrained and finetuned. For reference, we include models that were fine-tuned on the real datasets as well. It is important to note that the reported accuracies are not zero-shot for these two cases. First, we observe that training from scratch or finetuning on synthetic data using the CLIP loss results in substantially worse performance. We assess this behavior in detail in the next section. Based on this observation, we base our framework solely on feature distillation. We find that the resulting models outperform even the largest TinyCLIP models on three of the four datasets despite having 88% less trainable parameters. Moreover, they achieve comparable performance to the teacher models with a margin of 5% on three of the four datasets. Similarly, our model outperforms the largest TinyCLIP model in these cases. The larger performance gap on the Food dataset can likely be attributed to its larger test set size, which is also evident in the TinyCLIP models. When comparing to ViT-B/32 trained on DataComp-medium, which has been trained on a comparable number of images, even our purely pretrained models demonstrate significantly superior performance. Similarly, when pre-training a student using the CLIP loss on DataComp-medium for one epoch, the resulting zero-shot accuracies are far worse compared to the feature-based loss. This validates the data-efficiency of feature distillation for pre-training [12]. We report results for top-5 validation accuracy in the supplementary material.

4.3 Ablations

Based on the observation that models trained solely using feature distillation outperform those incorporating label-based losses when fine-tuning on synthetic data, we hypothesize that the utilization of labels during fine-tuning leads the model to learn spurious features, as well as domain-specific characteristics that

Table 1: The upper part summarizes the baseline CLIP and TinyCLIP. "Pre-train" denotes pre-training on a large but non domain-specific dataset. "Fine-tuned" contains the results where either synthetic data (zero-shot) or real data is used for fine-tuning the pre-trained model. Gray numbers indicate that performance is not zero-shot.

	Model	Loss	Training Datasets	Trainable Params. ImgEnc	Trainable Params. TxtEnc	#Samples Seen	Pets	Flowers	Cars	Food
CLIP	ViT-B/32	CLIP	DataComp-XL	86M	63M	12.8B	89.7	72.9	85.4	82.9
	ViT-B/32	CLIP	DataComp-medium	86M	63M	128M	43.1	29.7	28.0	41.7
	RN-50	CLIP	openai	86M	63M	32×400M	85.3	65.2	54.5	80.8
TinyCLIP	ViT-61M/32-29M	CLIP	LAION-400M	61M	29M	38×400M	87.3	64.7	79.1	73.4
	ViT-40M/32-19M	CLIP	LAION-400M	40M	19M	38×400M	84.4	61.0	74.2	71.4
	ViT-8M/16-3M	CLIP	YFCC-15M	8M	3M	50×15M	45.8	57.4	8.0	56.2
	RN-19M-19M	CLIP	LAION-400M	19M	19M	12×400M	81.0	56.4	70.1	66.7
Pre-train	TinyViT-11M	CLIP	DataComp-medium	11M	-	110M	10.4	4.2	5.4	4.7
	TinyViT-11M	$\mathcal{L}_2$	DataComp-medium	11M	-	110M	71.4	39.9	45.0	52.9
	TinyViT-11M	$\mathcal{L}_2$	DataComp-medium	11M	-	5× 110M	78.4	50.0	58.7	61.1
Fine-tuned	TinyViT-11M	$\mathcal{L}_2$	DataComp-medium	11M	-	110M				
		CLIP	+ Synthetic			+1M-9M	66.7	39.1	64.2	28.0
	TinyViT-11M	$\mathcal{L}_2$	DataComp-medium	11M	-	110M				
		$\mathcal{L}_2$	+ Synthetic			+1M-9M	**87.5**	**68.3**	**81.9**	**71.9**
	TinyViT-11M*	$\mathcal{L}_2$	DataComp-medium	11M	-	110M				
		CLIP	+ *Real* Train Images			+1M-7M	88.0	90.6	90.7	89.1
	TinyViT-11M*	$\mathcal{L}_2$	DataComp-medium	11M	-	110M				
		$\mathcal{L}_2$	+ *Real* Train Images			+1M-7M	88.7	68.4	83.8	83.0

distinguish between synthetic and natural images. Previously, it was observed that the using the CLIP loss hinders the class-incremental generalization of students distilled on real images [18]. We examine whether this observation transfers to the synthetic to real domain shift and potential spurious features. We highlight that the class-incremental setting is different from ours, as we use a fixed set of classes. To validate our hypothesis, we conducted three experiments on the pets dataset, deliberately introducing dedicated spurious features into synthetic and real images. Additionally, we evaluate students that we fine-tuned on either real or synthetic images on test sets from the respective other domain. Like that, we aim to provide further insights into the impact of label-based losses on the transferability of the models between synthetic and natural images. Subsequently, we employ larger and smaller students with two different architecture types to assess how the performance scales with the number of trainable parameters. Additionally, we report the performance of models trained from scratch.

Natural Images with Spurious Features. To investigate the impact of spurious features in the natural image domain, we add class-specific colored shapes to the images in the pets dataset. These shapes were added to each image in the training split, and examples can be seen in Figure 2. We then proceeded to fine-tune the pre-trained student models using the same hyperparameters employed during pre-training using these images. The test accuracies of the resulting models are shown in Table 2. We observe a slight decrease in performance on the test set without spurious features when the pre-trained students were fine-tuned with contrastive losses. This suggests that the students did not acquire any addi-

Table 2: Accuracy of the students trained on data with spurious features, introduced through adding colored shapes (real) or class-specific unicolor backgrounds (synthetic), on pets test set without spurious features. The spurious pets dataset used for testing features the same colored shapes but coupled with different classes (denoted by shuffled). Red indicates strong overfitting to trainsets with spurious features.

Trainset	Pets Testset	$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP	Teacher
Real	Real Testset	88.9	60.0	77.3	88.0	**90.0**	89.8
	Real Testset + Shuffled Spurious Features	88.0	48.7	51.7	88.0	**88.6**	88.7
	Real Trainset + Spurious Features	90.3	96.5	96.5	95.7	88.3	89.6
Synth.	Real Testset	**84.4**	24.3	31.0	81.2	35.6	89.7
	Synthetic Testset Without Spurious Features	90.9	53.6	61.7	**91.4**	66.7	93.9
	Synthetic Trainset + Spurious Features	94.2	100.0	100.0	99.3	100.0	93.3

tional class-specific features during fine-tuning. However, when evaluating these fine-tuned students on a test set of real images where the colored shapes were mixed between classes, we observe a significant degradation in performance. In contrast, the students trained with the $\mathcal{L}_2$ loss achieves accuracies comparable to the dataset without spurious features on both test sets. These findings highlight the robustness of the feature loss in mitigating the negative impact of spurious features in the natural image domain.

Synthetic Images with Spurious Features. To investigate whether the observed behavior on real images could be replicated using synthetic ones, we generated a synthetic dataset incorporating dedicated spurious features. Specifically, we sample images where pets are positioned against a solid-colored background, with each class assigned a distinct color. The results shown in Table 2 indicate that the performance of students trained with contrastive losses deteriorates when confronted with the presence of these spurious features. Figure 2 showcases instances of misclassifications. In contrast, the student trained with $\mathcal{L}_2$ loss exhibited a test accuracy of 84.4%, which is only 5% lower than the accuracy of the teacher despite the domain gap between real and synthetic images, as well as the presence of spurious features.

Generalization Between Synthetic and Real Images. In order to evaluate the ability of our models to generalize between synthetic and real images, we examine the performance of models that were fine-tuned on real training datasets when tested on independently generated synthetic datasets using the same methodology as the synthetic training sets. The results are presented in Table 3. The models fine-tuned with contrastive losses exhibited lower validation accuracy compared to the students trained with feature distillation. In contrast, for the models fine-tuned on synthetic data the reverse is true. Using the CLIP loss results in higher accuracy on the synthetic testset in comparison to feature distillation. This discrepancy suggests that the former models primarily learned domain-specific features of natural or synthetic images, thereby limiting their generalization capabilities between the two types.

Robustness to Common Corruptions. In order to evaluate the robustness of the models' learned representations against image perturbations, we conducted a

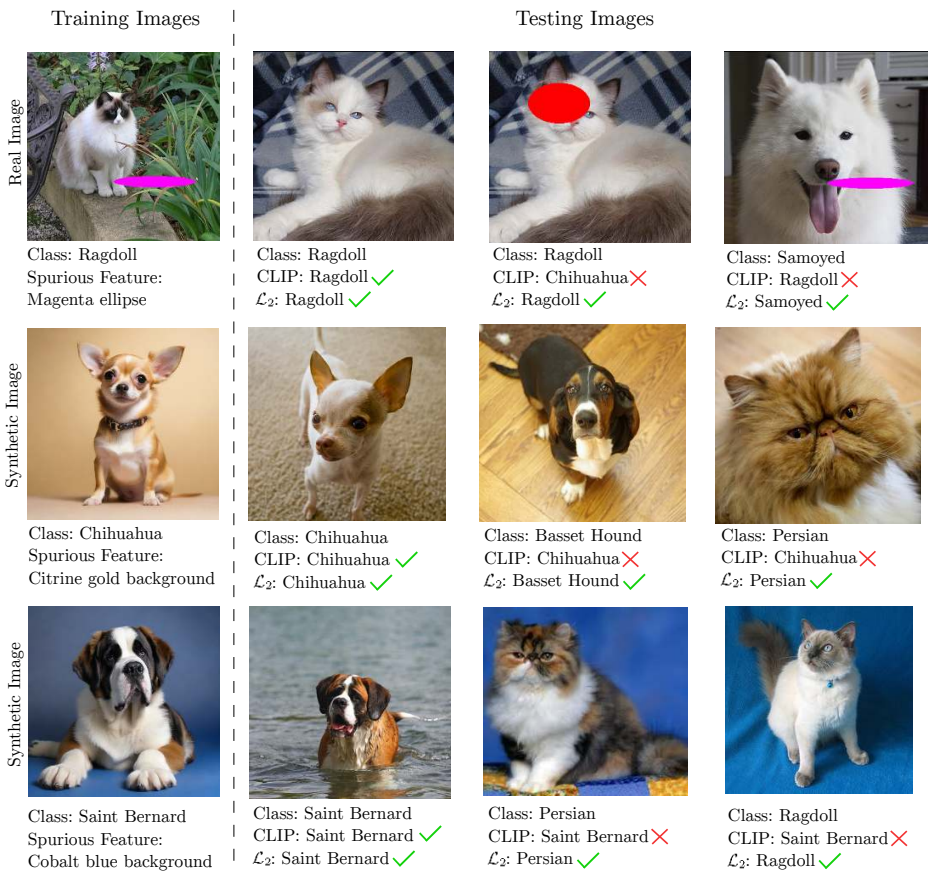

Fig. 2: Examples for misclassifications of the students fine-tuned with the CLIP loss. The first row corresponds to setting with natural images. The second and third row correspond to the student trained on synthetic images. All of the test examples are classified correctly by the teacher and the student trained with $\mathcal{L}_2$ loss.

comprehensive benchmark study. We assessed the performance of the classifiers on 15 common corruptions [13] at a fixed severity level of three, focusing on the pets dataset. The corresponding results are presented in Table 4. Our observations revealed that the students trained using the $\mathcal{L}_2$ feature loss demonstrate higher robustness to corruptions, regardless of whether they were trained on real or synthetic data. The distinction to the models fine-tuned with contrastive losses is particularly prominent when training on synthetic data, where the models fine-tuned with the CLIP loss even perform worse than the purely pre-trained model. These findings provide further evidence that contrastive losses lead to learning spurious and datatype-specific features, making the models less robust to disturbances caused by common corruptions. When evaluating on synthetic test data

Table 3: Classification accuracy of the fine-tuned models for pets on a synthetic test set which was independently but identically sampled as the synthetic train set.

Fine-Tuning Data	Test Data	Teacher Pre-Trained		$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Real	Synthetic	93.8	79.9	**91.7**	85.6	82.9	89.8	86.9
Synthetic	Synthetic	-	-	94.5	**97.9**	**97.7**	96.7	**97.8**

Table 4: Average performance of the models, which were fine-tuned on real data, on the pets testset under 15 common corruptions with severity 3.

Fine-Tuning Data	Test Data	Teacher Pre-Trained		$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Real	Real	78.2	52.4	73.6	65.3	64.8	**77.7**	66.7
Synthetic	Real	-	-	**66.2**	40.0	37.3	63.3	38.2
Real	Synthetic	93.6	73.5	93.8	94.5	94.2	**95.2**	93.2
Synthetic	Synthetic	-	-	90.8	83.5	83.4	**91.2**	85.4

with common corruptions, we observe that the performance drops are lower and the characteristics of the synthetic data are less disturbed by corruptions.

Student Model Size. Alongside the results of a TinyViT-11M in Table 1, we report the zero-shot performance of five additional students after pre-training and fine-tuning using feature distillation in Figure 3. As expected, there is a general trend of improved performance with increasing model size. Yet, we observe that the effectiveness of zero-shot fine-tuning is more pronounced for smaller models compared to larger ones: after fine-tuning, the difference in performance between the largest and smallest models is around 10% to 15%, while after purely pre-training it is as high as 30%. These findings highlight that zero-shot distillation is particularly effective for smaller models since it allows to adapt them to the target domain, without requiring real in-domain data.

Training From Scratch. We investigate the accuracy of models trained from scratch, both on real and synthetic images in Fig. 4. On real data, using a sum of the $\mathcal{L}_2$ and a contrastive loss indeed results in the best performance as observed in CLIP-KD [50]. The low accuracy of the feature distilled student on the flowers dataset can be attributed to the small training set with only 10 images per class. On synthetic training data, the $\mathcal{L}_2$ loss consistently performs best while the CLIP loss or the MP loss results are substantially worse. Combing the $\mathcal{L}_2$ with one of the contrastive losses does not yield any performance gains in contrast to purely feature-based distillation. This is in line with the behavior observed during fine-tuning in Tab. 1.

5 Conclusion

In this work, we introduced a framework for zero-shot distillation of small CLIP image encoders based on synthetic images. We make the surprising observation that contrastive losses are a potentially detrimental factor for generalization capabilities of models between synthetic and real data, due to the exploitation

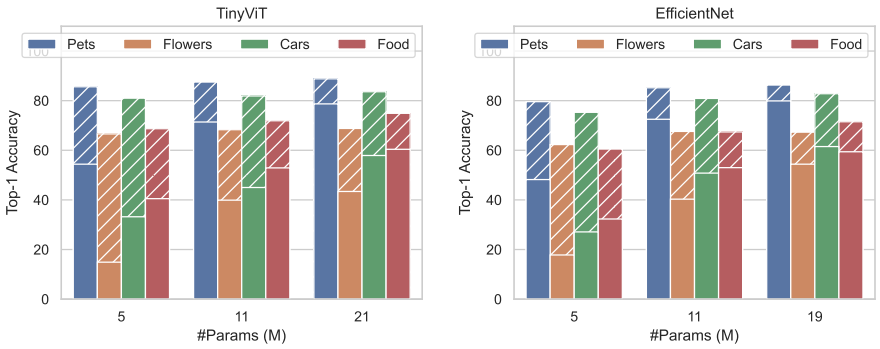

Fig. 3: Zero-shot classification performance of the students after pre-training on DataComp medium for one epoch and after finetuning on the synthetic datasets for 96 epochs (hatched). All experiments were performed only using feature distillation.

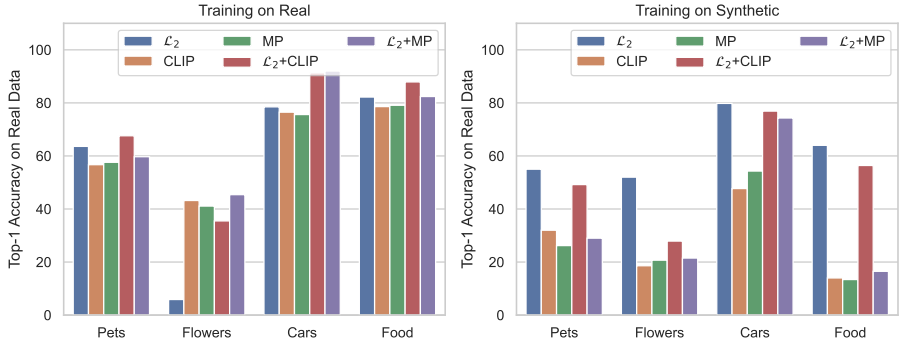

Fig. 4: Classification performance of the models after training from scratch for 96 epochs on either real or synthetic images.

of spurious features. By employing a pure image feature-based distillation loss, we successfully mitigate this limitation. As a result, we are able to train models that surpass the current state-of-the-art for zero-shot CLIP distillation, while featuring fewer parameters and not using labeled target domain images.

Limitations and Future Work. The presented results are based on a ViT-B/32 teacher; future work could explore the benefit of larger teachers for constant-sized students. Furthermore, the potential of our small-scale image encoders beyond zero-shot settings could be explored, for instance in architectures that use CLIP image encoders such as BLIP-2 [17] or LLava [21–23]. Moreover, the current framework is limited to classification tasks. To broaden its applicability, future research could extend the framework to encompass other computer vision tasks such as object detection or image segmentation.

Acknowledgements. We would like to thank Nicole Finnie for helpful discussion on pre-training CLIP models. We also thank the European Laboratory for Learning and Intelligent Systems (ELLIS) for supporting Niclas Popp.

References

1. Ahmed, B.S., Zamli, K.Z., Afzal, W., Bures, M.: Constrained interaction testing: A systematic literature study. IEEE Access **5**, 25706–25730 (2017) 6, 19
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. Transactions on Machine Learning Research (2023) 2, 4, 5, 21
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: ECCV (2014) 3, 8
4. da Costa, V.G.T., Dall'Asen, N., Wang, Y., Sebe, N., Ricci, E.: Diversified in-domain synthesis with efficient fine-tuning for few-shot classification. arXiv:2312.03046 (2023) 4, 6
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009) 4
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 8
7. Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marten, R., Wortsman, M., Ghosh, D., Zhang, J., Orgad, E., Entezari, R., Daras, G., Pratt, S., Ramanujan, V., Bitton, Y., Marathe, K., Mussmann, S., Vencu, R., Cherti, M., Krishna, R., Koh, P.W., Saukh, O., Ratner, A., Song, S., Hajishirzi, H., Farhadi, A., Beaumont, R., Oh, S., Dimakis, A., Jitsev, J., Carmon, Y., Shankar, V., Schmidt, L.: Datacomp: In search of the next generation of multimodal datasets. arXiv:2304.14108 (2023) 2, 5, 7, 8
8. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J.: Vision-language pre-training: Basics, recent advances, and future trends. Found. Trends. Comput. Graph. Vis. **14**(3–4), 163–352 (dec 2022) 5
9. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. Int. J. Comput. Vision **129**(6), 1789–1819 (jun 2021) 6
10. Goyal, S., Kumar, A., Garg, S., Kolter, Z., Raghunathan, A.: Finetune like you pretrain: Improved finetuning of zero-shot vision models. In: CVPR (2023) 7
11. Hammoud, H.A.A.K., Itani, H., Pizzati, F., Torr, P., Bibi, A., Ghanem, B.: Synthclip: Are we ready for a fully synthetic clip training? arXiv:2402.01832 (2024) 4, 5, 6, 21, 24
12. He, R., Sun, S., Yang, J., Bai, S., Qi, X.: Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability. In: CVPR (2022) 4, 5, 9
13. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019) 12
14. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv:1503.02531 (2015) 3, 6, 21, 27
15. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling BERT for natural language understanding. In: Cohn, T., He, Y., Liu, Y. (eds.) EMNLP (2020) 3

16. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013) 3, 8
17. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023) 14
18. Li, X., Fang, Y., Liu, M., Ling, Z., Tu, Z., Su, H.: Distilling large vision-language model with out-of-distribution generalizability. In: ICCV. pp. 2492–2503 (October 2023) 4, 10, 25
19. Li, Z., Li, Y., Zhao, P., Song, R., Li, X., Yang, J.: Is synthetic data from diffusion models ready for knowledge distillation? arXiv:2305.12954 (2023) 4, 21
20. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In: NeurIPS (2022) 5
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv:2310.03744 (2023) 14
22. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 14
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 14
24. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 8, 27
25. Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J., Zhao, H.: Lcm-lora: A universal stable-diffusion acceleration module. arXiv:2311.05556 (2023) 8, 19
26. Metzen, J.H., Hutmacher, R., Hua, N.G., Boreiko, V., Zhang, D.: Identification of systematic errors of image classifiers on rare subgroups. In: ICCV. pp. 5064–5073 (October 2023) 19
27. Mirzadeh, S.I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: AAAI (2020) 3
28. Nayak, G.K., Mopuri, K.R., Shaj, V., Babu, R.V., Chakraborty, A.: Zero-shot knowledge distillation in deep networks. In: ICML. pp. 4743–4751 (2019) 5
29. Nie, C., Leung, H.: A survey of combinatorial testing. ACM Comput. Surv. **43**, 11 (02 2011) 6, 19
30. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (2008) 3, 8
31. Ojha, U., Li, Y., Rajan, A.S., Liang, Y., Lee, Y.J.: What knowledge gets distilled in knowledge distillation? (2023) 3, 4
32. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023) 1
33. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: CVPR (2012) 3, 8
34. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv:2307.01952 (2023) 2, 8, 19

35. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021) 1, 2, 5, 7, 8, 23, 24
36. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:1910.01108 (2020) 3
37. Sariyildiz, M.B., Karteek, A., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR (2022) 2, 6, 21
38. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv:2111.02114 (2021) 5, 9
39. Stanton, S.D., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? In: NeurIPS (2021) 4
40. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: ICML (2019) 8
41. Thomee, B., Elizalde, B., Shamma, D., Ni, K., Friedland, G., Poland, D., Borth, D., Li, L.J.: Yfcc100m: the new data in multimedia research. Communications of the ACM **59**, 64–73 (01 2016) 9
42. Tian, Y., Fan, L., Chen, K., Katabi, D., Krishnan, D., Isola, P.: Learning vision from models rivals learning vision from data. arXiv:2312.17742 (2023) 4, 5, 21, 24
43. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. In: NeurIPS (2023) 5, 7, 21, 24
44. Tian, Y., Krishnan, D., Isola, P.: Contrastive multiview coding. p. 776–794. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-58621-8_45, https://doi.org/10.1007/978-3-030-58621-8_45 23
45. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 (2023) 8, 19
46. Vasu, P.K.A., Pouransari, H., Faghri, F., Vemulapalli, R., Tuzel, O.: Mobileclip: Fast image-text models through multi-modal reinforced training. arXiv:2311.17049 (2023) 4, 9, 21
47. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: NeurIPS. pp. 10506–10518 (2019) 25, 26
48. Wu, K., Peng, H., Zhou, Z., Xiao, B., Liu, M., Yuan, L., Xuan, H., Valenzuela, M., Chen, X.S., Wang, X., Chao, H., Hu, H.: TinyCLIP: CLIP distillation via affinity mimicking and weight inheritance. In: ICCV (2023) 2, 3, 5, 21
49. Wu, K., Zhang, J., Peng, H., Liu, M., Xiao, B., Fu, J., Yuan, L.: Tinyvit: Fast pretraining distillation for small vision transformers. In: ECCV (2022) 3, 27
50. Yang, C., An, Z., Huang, L., Bi, J., Yu, X., Yang, H., Xu, Y.: CLIP-KD: An empirical study of distilling clip models. arXiv:2307.12732 (2023) 3, 9, 13

51. Yu, Z., Zhu, C., Culatana, S., Krishnamoorthi, R., Xiao, F., Lee, Y.J.: Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. arXiv:2312.02253 (2023) 2, 4, 5, 6, 21
52. Zhang, T., Wang, Z., Huang, J., Tasnim, M.M., Shi, W.: A survey of diffusion based image generation models: Issues and their solutions. arXiv:2308.13142 (2023) 19
53. Zhou, A., Wang, J., Wang, Y.X., Wang, H.: Distilling out-of-distribution robustness from vision-language foundation models. In: NeurIPS (2023), https://openreview.net/forum?id=iwp3H8uSeK 25

A Supplementary Material

A.1 Details of the Synthetic Data Generation

In this section, we provide further details on the synthetic data generation and the diversification process. As mentioned in Section 3.3, the prompts used to synthesize the images are based on the classnames and additional information given by an LLM. For each class, we ask the language model to provide information with respect to four contextual dimensions as well as a superclass. The contextual dimensions are dataset specific and are summarized in Table 5. Figure 6 shows a concrete example for a class from the pets dataset. For each of the contextual dimensions we collect 15 or 30 options from Llama 2 7B fine-tuned for chats [45]. The larger number of options for the food dataset is used to accommodate its larger test set. In Table 6 we summarize the sizes of the real target datasets. Instead of using all possible combinations of options for the contextual dimensions to generate the prompts, we use combinatorial testing [1, 29]. This approach is inspired by a recent work on systematic error identification [26]. It reduces the number of images per class while ensuring that the prompts systematically cover the diversity contained in the answers from the LLM. For example, in case of 15 options per contextual dimension, this results in 265 images per class instead of 50625. The prompts are a comma separated list of the selected options, which are weighted to accommodate for contextual dimensions that are more or less important for certain domains. These weighted prompts are then used as input for a diffusion model. Specifically, we use Stable Diffusion XL [34] LCM LoRA [25]. To ensure sufficient image quality and diversity we employ a guidance scale of 0.5 and 6 inference steps. Further example images can be found in Figure 5. These also showcase some of the known problems with diffusion models such as parts of the prompts which are missing in the image [52] as in the first example for the food dataset.

Table 5: Contextual dimensions and prompt weights for the diversified data generation

Dataset	Weight of Classname	Contextual Dimensions and Weights in Braket	Options per Dimension	Images per Class
Pets	1.5	superclass (1.2), locations, position, daytime, camera angle	15	265
Flowers	1.2	superclass, color, locations, daytime (0.1), camera angle (0.1)	15	265
Cars	1.0	superclass, locations, color, daytime, camera angle	15	265
Food	1.2	superclass, locations, way of serving (1.5), daytime(0.1), camera angle(0.1)	30	1011

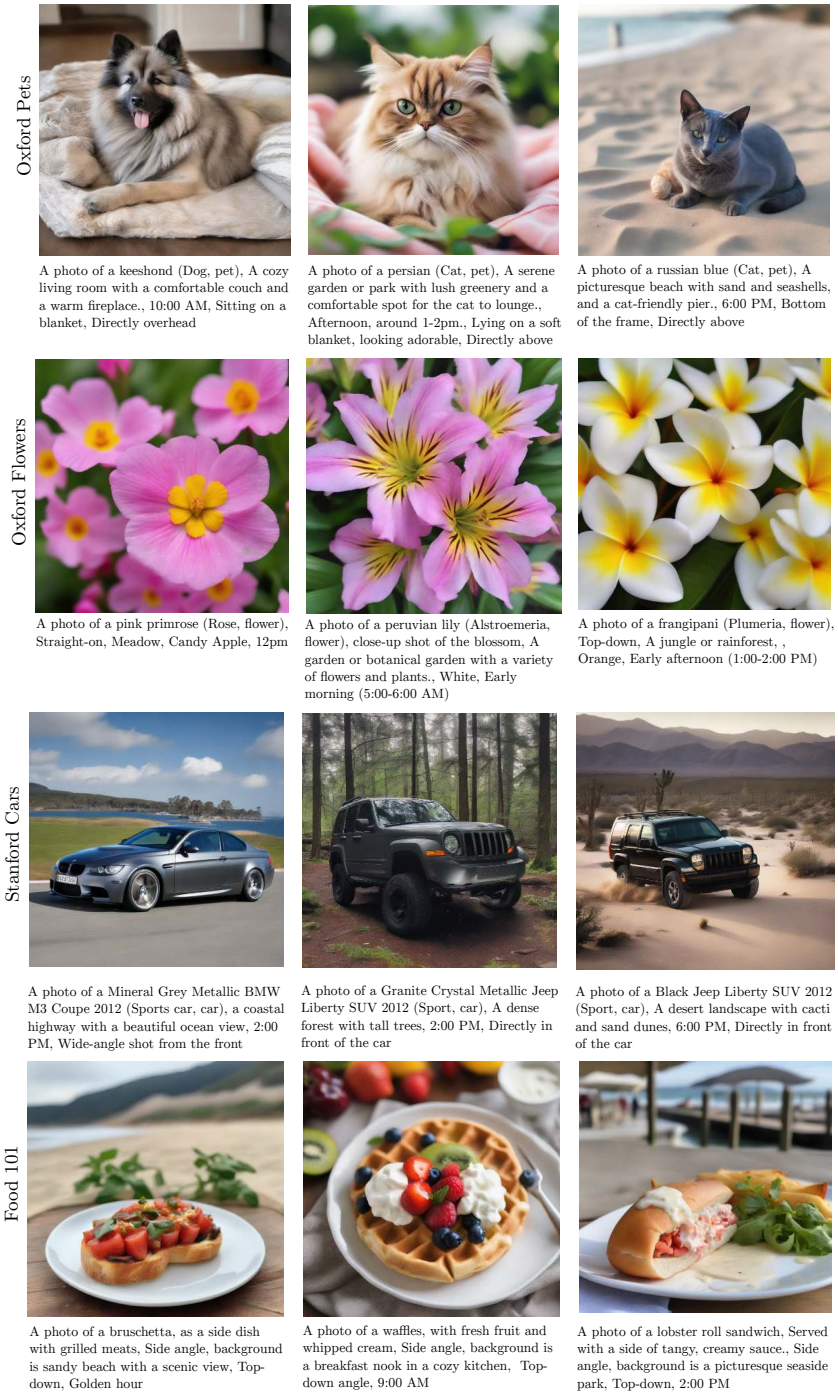

Fig. 5: Examples for synthetic images from all four datasets together with the prompts used to generate them.

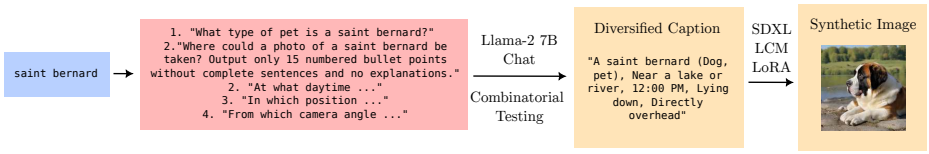

Fig. 6: The figure illustrates the process for generating synthetic images for the example class "saint bernard" from the pets dataset.

Table 6: Overview over the size of the real target datasets

Dataset	#classes	#training images	#test images
Pets	37	3680	3669
Flowers	102	1020	6149
Cars	196	8144	8041
Food	101	75750	25250

A.2 Overview over (Zero-Shot) Baselines

To complement the framework presented in Section 4, we situate existing baselines with respect to the discussed components in Table 7. We state whether pre-training and/or fine-tuning is used on synthetic or natural images or both. Data diversification describes if the approach uses prompt or images diversification for the synthetic data. We note that our framework unifies existing pipelines.

A.3 Top-5 Test Accuracies

To complement the results in Section 4, we provide the top-5 accuracies of the pre-trained and fine-tuned models. Figure 7 visualizes the results. The trends mirror the observations from the Top-1 accuracy with an even smaller gap to the teacher.

Table 7: Comparison of existing training approaches to our zero-shot distillation framework. Similar to our framework, the DM-KD pipeline only relies on synthetic images for training on specific target datasets. However, the teacher was trained on the real images, which is not possible in a zero-shot setting. This is symbolized by *.

Name	Evaluation Metric	Pre-Training	Fine-Tuning	Natural Images	Synthetic Images	Data Diversification	Loss
StableRep [43]	Linear probe, few-shot	✓			✓		MP
SynCLR [42]	Linear probe	✓			✓	✓	MP
SynthCLIP [11]	Linear probe, few-shot	✓			✓	✓	CLIP
Fake it till you make it [37]	Zero-Shot Acc.	✓			✓		Cross-Entropy
Diversify don't finetune [51]	Accuracy	✓		✓	✓	✓	Custom
[2]	Accuracy	✓		✓	✓		Cross-Entropy
DM-KD [19]	Accuracy*		✓		✓		KD [14]
TinyCLIP [48]	Zero-Shot Acc.	✓		✓			CLIP, Affinity Mapping
MobileCLIP [46]	Zero-Shot Acc.	✓		✓			CLIP, Affinity Mapping
Zero-Shot Distillation (Ours)	Zero-Shot Acc.	✓	✓	✓	✓	✓	$\mathcal{L}_2^{feature}$

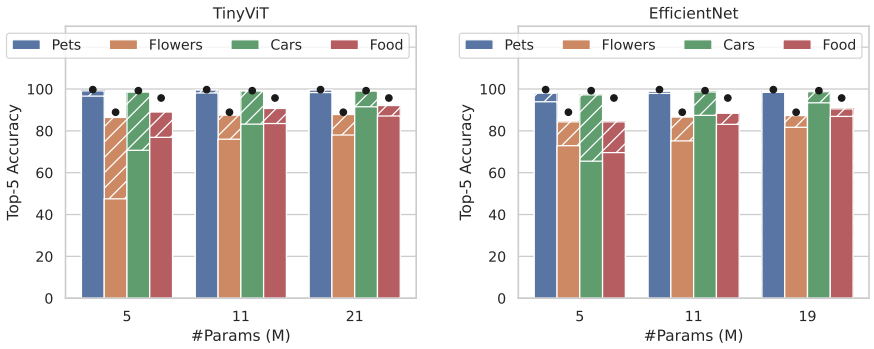

Fig. 7: Top-5 accuracy of the models after pre-training on DataComp medium for one epoch and after fine-tuning on the synthetic datasets for 96 epochs (hatched). The black dots indicate the top-5 accuracy of the teacher.

Table 8: Accuracy of the fine-tuned models on the datasets they were trained on.

Trainset	Loss	Pets Top-1	Pets Top-5	Flowers Top-1	Flowers Top-5	Cars Top-1	Cars Top-5	Food Top-1	Food Top-5
Real	$\mathcal{L}_2$	87.7	99.0	70.3	89.3	83.8	99.1	79.3	93.7
	CLIP	89.7	98.5	97.3	100.0	90.6	100.0	97.6	99.2
	MP	89.1	98.5	97.5	100.0	90.7	100.0	97.5	99.2
	$\mathcal{L}_2$+CLIP	91.2	99.9	90.6	98.0	92.2	100.0	90.8	98.3
	$\mathcal{L}_2$+MP	100.0	100.0	97.7	100.0	92.8	100.0	98.5	99.6
Synthetic	$\mathcal{L}_2$	95.3	100.0	68.0	90.1	75.9	95.7	84.5	96.3
	CLIP	100.0	100.0	97.8	98.9	90.0	99.2	99.7	100.0
	MP	99.9	100.0	99.5	100.00	87.6	98.4	99.7	100.0
	$\mathcal{L}_2$+CLIP	97.9	100.0	85.8	96.9	91.0	99.7	93.2	98.8
	$\mathcal{L}_2$+MP	100.0	100.0	99.6	100.0	94.7	100.0	99.9	100.0

A.4 Training Accuracies

In addition to the test accuracies reported in the main paper, we report the training accuracies of the fine-tuned models in Table 8. We observe that the models trained with a contrastive loss achieve higher training accuracy than the feature distilled students. This holds in particular on synthetic data. In combination with the results from Section 4, this underlines our hypothesis that contrastive losses lead to learning domain specific or spurious features over actual class or object specific features.

A.5 Contrastive Image Loss

As an alternative to the feature-based $\mathcal{L}_2$ loss, we test a contrastive loss that is purely based on the image feature of the student and teacher. Using the notation

from Section 3.4, it is defined as

$$\mathcal{L}_{\text{contrastive}}^{\text{image}} = \sum_{i=1}^{N} -\log \frac{\exp(\langle \mathbf{I}_i^S, \mathbf{I}_i^T \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{I}_i^S, \mathbf{I}_k^T \rangle / \tau)} \tag{1}$$

where τ denotes a learnable temperature parameter. We use this loss both for pre-training and fine-tuning a TinyViT 11M with the same setup as for the $\mathcal{L}_2$ loss in Section 4. The results are reported in Table 9. We observe that for pre-training, the contrastive image loss results in better performance in comparison to the $\mathcal{L}_2$ loss, while for fine-tuning it is the other way around. Yet, the contrastive image loss still clearly outperforms the CLIP loss when fine-tuning on synthetic data. This validates our observation from Section 4 that using a purely feature based loss improves the generalization between synthetic and real data.

A.6 ImageNet-100 Fine-Tuning

As an addition to the domain specific datasets discussed in the main paper, we fine-tune the models on ImageNet-100 [44] which is a subset of ImageNet-1k consisting of 100 classes with various objects that do not necessarily belong to a similar domain. To generate synthetic images, we use the same setup as for the datasets in Section 4 with prompt diversification and 30 options per contextual dimension. This equals 1011 images per class. The contextual dimensions are the same as for the cars dataset but without any prompt weighting. Instead of using the simple zero-shot prompts "a photo of a ..., which is a type of ...", we use the prompt ensembles consisting of 79 templates proposed by Radford et al. [35]. These prompt templates are also used for the diversified prompts by starting

Table 9: Accuracy of the models that were pre-trained and fine-tuned using distillation with the contrastive image loss. The differences to $\mathcal{L}_2$ feature distillation and training using the CLIP loss are shown in gray.

Trainset	Training	Loss	Pets	Flowers	Cars	Food
Real	Pre-Trained	Contrastive Image	72.8	39.6	46.5	54.5
		Difference to $\mathcal{L}_2$	+1.4	+0.5	+1.5	+1.4
		Difference to Contrastive Image-Text (CLIP)	+52.4	+35.4	+41.1	+49.8
	Fine-Tuned	Contrastive Image	83.9	64.9	81.4	80.4
		Difference to $\mathcal{L}_2$	-4.8	-3.5	-2.4	-2.6
		Difference to Contrastive Image-Text (CLIP)	-5.5	-25.7	-9.3	-2.6
Synth.	Fine-Tuned	Contrastive Image	80.5	57.7	79.3	68.8
		Difference to $\mathcal{L}_2$	-7.0	-10.6	-2.6	-3.1
		Difference to Contrastive Image-Text (CLIP)	+13.8	+18.6	+15.1	+40.8

Table 10: Zero-Shot accuracy of a TinyViT 11M pre-trained for one epoch on DataComp medium and fine-tuned on ImageNet-100 using either real or synthetic data. Using the CLIP loss with synthetic data decreases the performance in contrast to pure pre-training. Fine-tuning through $\mathcal{L}_2$ feature distillation yields a slight improvement.

Training Data	Teacher	Pre-Trained	$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Real	86.1	74.3	81.8	87.0	87.7	87.2	**89.3**
Synthetic	-	-	74.0	53.2	52.4	**74.2**	68.2

with a randomly chosen one before listing the classname and the options for the contextual dimensions. The results are shown in Table 10. For the contrastive losses, we observe the same trends as for the domain specific datasets. When fine-tuning on real data, the resulting test performance exceeds the teacher while fine-tuning with synthetic data deteriorates the accuracy in contrast to pure pre-training. For the $\mathcal{L}_2$ loss this is not the case. Yet, fine-tuning with feature distillation yields almost the same zero-shot accuracy as pure pre-training. This is different to the domain specific datasets where we could observe a consistent improvement. This is likely due to the larger diversity in the real test images which is not sufficiently captured by the synthetic training data.

A.7 Linear Probing

To evaluate the linear probe accuracy of the teacher as well as the TinyViT 11M models pre-trained and fine-tuned on the pets dataset, we fit a linear classifier based on the unnormalized image features after the projection head. The classifier is fitted either using synthetic or real data, where only the case of synthetic data corresponds to the zero-shot setting. The hyperparameter sweeps for the regularization are performed on a validation split as in the original CLIP paper [35]. The results are shown in Table 11. For the models fine-tuned on synthetic data, the performance is 8 to 10 % worse when probing with synthetic data in comparison to fitting the linear classification head with real data. This highlights that using linear probing based on real data improves the linear classification accuracy by a substantial margin. In contrast, the true zero-shot linear classifiers, where the classification head is fitted using synthetic data, perform comparable to or worse than pure zero-shot classification using the similarity between the image and prompt embeddings. In contrast to our framework, previous works [11,42,43] mainly focus on the linear accuracy where the classification head is fitted with real data instead of targeting the true zero-shot setting without any real data which is the more difficult task to accomplish.

A.8 Performance Under Domain Shift

In addition to the synthetic-real domain gap, we investigate the accuracy of the fine-tuned models under dedicated domain shift. This setting is comparable to

Table 11: Linear probe accuracy of the teacher as well as TinyViT 11Ms models pre-trained for one epoch on DataComp medium and fine-tuned on pets. The linear accuracy of the models that were fine-tuned with synthetic data is increased substantially by probing with real data compared to probing with synthetic data.

Fine-Tuning Data	Probing Data	Teacher Pre-Trained		$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Real	Real	90.4	81.6	89.8	88.3	88.9	92.1	90.2
Real	Synthetic	84.0	71.8	82.1	87.7	88.3	87.7	88.9
Synthetic	Real	-	-	89.6	73.7	72.6	90.1	75.3
Synthetic	Synthetic	-	-	80.9	64.5	65.0	82.8	65.8

Table 12: Accuracy of the teacher as well as TinyViT 11Ms models pre-trained for one epoch on DataComp medium and fine-tuned on synthically generated pets test data which comprises of sketches. The feature-distilled student exhibit a substantially smaller drop in performance compared to the contrastively fine-tuned models.

Fine-Tuning Data	Teacher Pre-Trained		$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Real	89.3	71.3	83.3	70.4	72.6	83.8	73.0
Difference to real test data	-0.4	-7.1	-5.4	-17.6	-16.4	-13.9	-17.6
Synthetic	-	-	86.9	76.0	75.2	87.9	81.9
Difference to synthetic test data			-7.6	-21.5	-22.5	-8.8	-15.9

the experiments performed by Zhou et al. [53]. However, our setup requires no adversarial training. We sample a dataset of images that utilizes the prompt template "a sketch of a ..., which is a type of ..." instead of "a photo of a ..., which is a type of ..." to generate sketches rather than photorealistic images. The remainders of the prompts are identical to those of the independent synthetic test set. Figure 8 illustrates example images from this dataset. For zero-shot classification, we continue to employ "a photo of a ..., which is a type of ..." to simulate unknown domain shift. The results are presented in Table 12. When comparing to the testsets without dedicated domain shifts, we observe that the feature distilled student achieves the lowest decrease in performance, substantially outperforming the models that were fine-tuned using purely contrastive losses. Interestingly, the drop in accuracy is larger when using synthetic in comparison to real data for fine-tuning where the test set feature the domain shift from photos to sketches but not from synthetic to real images. Additionally, we test the models fine-tuned on ImageNet-100 on the corresponding classe of ImageNet Sketch [47]. The results are shown in Table 13. In this case, the students distilled with the feature loss are again more robust against the shift to sketches. Our observations on domain shift mirror the behavior in the class-incremental setting [18]. There, the models are trained on a subset of the real training images that contains only a fraction of the overall classes and are evaluated on the set of remaining classes. In our case, however, we use a fixed number of classes.

A sketch of a keeshond (Dog, pet), A cozy living room with a comfortable couch and a warm fi replace., 10:00 AM, Sitting on a blanket, Directly overhead

A sketch of a persian (Cat, pet), A serene garden or park with lush greenery and a comfortable spot for the cat to lounge., Afternoon, around 1-2pm., Lying on a soft blanket, looking adorable, Directly above

A sketch of a russian blue (Cat, pet), A picturesque beach with sand and seashells, and a cat-friendly pier., 6:00 PM, Bottom of the frame, Directly above

Fig. 8: Examples from the synthetic test set which comprises of sketches.

Table 13: Zero-Shot accuracy of the models fine-tuned on ImageNet-100 when tested on the 100 corresponding classes in ImageNet Sketch [47]. The students distilled using feature distillation exhibit better performance as well as lower degration in comparison to the accuracy on standard ImageNet-100

Fine-Tuning Data	Teacher	Pre-Trained	$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Real	76.3	54.1	57.5	49.4	51.0	**60.6**	56.8
Difference to real test data	-14.1	-20.2	-24.3	-38.0	-36.7	-27.2	-32.5
Synthetic	-	-	**56.3**	34.0	25.6	55.7	49.8
Difference to real test data			-17.7	-19.2	-26.8	-18.5	-18.4

A.9 Simple Prompts

To assess the influence of diverse prompts on image generation, we utilized zero-shot prompts "a photo of a ..., which is a type of ..." to generate images instead of diversified prompts from a LLM. We sample a synthetic pets dataset with the same number of images per class as in the diversified case. Example images are shown in Figure 9. The diversity of the images decreases, especially with regard to the camera angle, as almost all images show only a frontal shot of the animals with the focus on the face. Furthermore, the variety of backgrounds decreases. We fine-tune a TinyViT 11M model on this dataset and observed the results in Table 14. The accuracy of feature distilled student exhibits only a small decrease in performance, while the models which were fine-tuned with contrastive losses significantly decreased. These findings indicate that the students distilled with the $\mathcal{L}_2$ feature loss can deal better with a lack of diversity during fine-tuning.

A.10 Linear Classification Head Instead Of CLIP Architecture

Instead of using the CLIP architecture, we train and distill TinyViT 11M model with a linear classification head on the pets dataset for comparison. We either

Fig. 9: Examples from the synthetic dataset set which was generated using the zero-shot prompts "a photo of ...". The resulting images feature less diversity in comparison to the diversified prompts generated by an LLM depicted in Figure 5.

Table 14: Accuracy of the TinyViT 11Ms models pre-trained for one epoch on DataComp medium and fine-tuned on synthetically generated pets test data was sampled with the zero shot prompts "a phot of ..." instead of diverse prompts from a LLM. The performance of feature-distilled students degrade considerably less from a lack of diversity than the contrastively fine-tuned models.

Fine-Tuning Data	Test Data	$\mathcal{L}_2$	CLIP	MP	$\mathcal{L}_2$+CLIP	$\mathcal{L}_2$+MP
Synthetic (simple)	Real	85.9	40.7	45.2	81.8	49.1
Difference to synthetic (diversified)		-1.6	-26.0	-21.3	-5.4	-19.0
Synthetic (simple)	Synthetic (diversified)	93.0	86.6	85.0	93.8	87.6
Difference to synthetic (diversified)		-1.5	-10.9	-12.7	-2.9	-10.2

train from scratch or fine-tune models with pre-trained weights from ImageNet-22k (with the exception of the linear classification head which is always randomly initialized). As most of the classes from the pets dataset are contained in ImageNet-22k, the latter does not correspond to a strict zero-shot setting even when fine-tuning with synthetic data. To train the models, we optimize the standard cross-entropy loss as well as a sum of cross-entropy loss and the original knowledge distillation loss of Hinton et al. [14]. We use the AdamW optimizer [24] with no weight decay and the learning rate is set to 5×10^{-4} which is the same as used by Wu et al. [49] for fine-tuning. First, we observe that the drop in performance when fine-tuning with synthetic data in comparison to real data is similar to the CLIP models based on contrastive losses. Additionally, the performance of the model with classification head is worse compared to the CLIP models when trained from scratch. In contrast, the classifiers pretrained on ImageNet-22k and fine-tuned on the real training data achieve the best performance overall. This can presumably be attributed to the fact that most of the classes are already included in ImageNet-22k which was used for pre-training. Using knowledge distillation for the models with classification head only has a minor effect.

Table 15: Accuracy of the models that feature a linear classification head instead of a CLIP architecture. Training was performed on the synthetic and real pets datasets for 96 epochs using the cross-entropy loss (CE) with or without knowledge distillation (KL).

Fine-Tuning Data	Pre-Training Data	CE	CE+KL
Real	-	47.8	47.2
Synthetic	-	26.9	29.9
Real	ImageNet-22k	91.2	91.6
Synthetic	ImageNet-22k	71.3	67.4