Low-Resource Vision Challenges for Foundation Models

Yunhua Zhang[1] Hazel Doughty[2] Cees G. M. Snoek[1]
[1]University of Amsterdam [2]Leiden University

Abstract

Low-resource settings are well-established in natural language processing, where many languages lack sufficient data for deep learning at scale. However, low-resource problems are under-explored in computer vision. In this paper, we address this gap and explore the challenges of low-resource image tasks with vision foundation models. We first collect a benchmark of genuinely low-resource image data, covering historic maps, circuit diagrams, and mechanical drawings. These low-resource settings all share three challenges: data scarcity, fine-grained differences, and the distribution shift from natural images to the specialized domain of interest. While existing foundation models have shown impressive generalizability, we find they cannot transfer well to our low-resource tasks. To begin to tackle the challenges of low-resource vision, we introduce one simple baseline per challenge. Specifically, we i) enlarge the data space by generative models, ii) adopt the best sub-kernels to encode local regions for fine-grained difference discovery and iii) learn attention for specialized domains. Experiments on our three low-resource tasks demonstrate our proposals already provide a better baseline than transfer learning, data augmentation, and fine-grained methods. This highlights the unique characteristics and challenges of low-resource vision for foundation models that warrant further investigation. Project page: `https://xiaobai1217.github.io/Low-Resource-Vision/`.

1. Introduction

Many have studied low-resource natural language processing [8, 23, 32, 57, 89], in which the target languages are less common and data is scarce. In computer vision, numerous works have explored effective learning methods for limited labeled data scenarios, *e.g.*, by meta-learning [35, 37], few-shot learning [53, 67], or generative modeling [24, 85]. Albeit successful, they focus on high-resource image domains, where thousands of images from the same domain are available, even though each class may only have a few samples for model learning. Different from existing works handling data scarcity, we investigate low-resource settings for computer vision where data is truly scarce (see Figure 1).

By collecting a benchmark of low-resource vision tasks we are able to study the combination of challenges unique to this area. First, data is severely limited with only a few hundred examples available for model learning. Second, vision tasks that are low-resource also tend to be highly specialized, meaning differences between different images are extremely subtle and fine-grained. Finally, the limited examples and specialized nature of the task means the domain is incredibly different from more common natural images available in bulk. While these challenges are often studied in isolation [34, 68] or even pairs [12, 46], the combination of all three is unique to low-resource vision and demands solutions outside of the scope of current models.

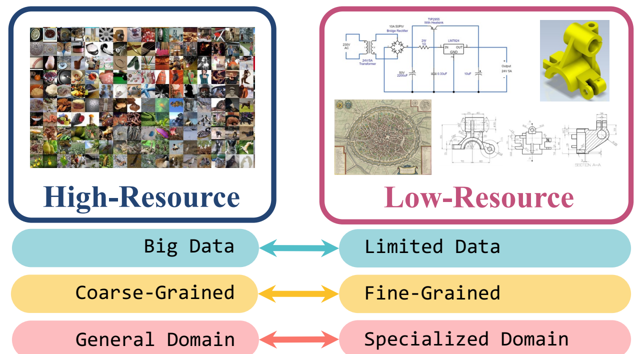

Figure 1. **High-Resource vs Low-Resource Vision**. High-resource vision focuses on images that can be collected at scale, have coarse-grained classes, and come from a general domain. We instead focus on low-resource vision tasks where data is scarce, has fine-grained differences, and comes from highly specialized domains.

An intuitive way to handle low-resource vision is leveraging the strong image representations from foundation models, which have progressed at a tremendous pace in recent years [27, 39, 44, 49, 58]. They show promising zero-shot performance on various downstream vision tasks and provide representations with generalization and transfer capabilities. Thus, they are a natural solution to low-resource vision. However, we find that current foundation models [27, 49, 50, 58] struggle to generalize to the specialized domains of low-resource vision tasks. We also find that existing transfer learning techniques struggle to adapt with the very limited amount of data available. Thus, we propose several adaptation baselines to begin to tackle the challenges of low-resource vision, with the ambition to inspire future work in this area.

As our main contribution, we study the challenges of low-resource vision and collect a low-resource image benchmark. Specifically, our benchmark covers circuit diagrams, historic maps, and mechanical drawings. We find the challenges of low-resource vision are a lack of training data, fine-grained differences, and domain shift from natural images to specialized domains. From our analysis, we discover foundation models struggle to recognize and retrieve low-resource images as do existing transfer learning methods. Thus, we introduce three simple baselines to mitigate each difficulty. Specifically, we finetune foundation models using diverse data produced by generative models to cope with data scarcity, we discover fine-grained details by focusing on local patterns via selected sub-kernels, and we learn attention for specialized domains to combat the distribution shift. Experiments demonstrate the challenges of our low-resource benchmark for existing transfer learning, data augmentation, and fine-grained methods as well as the advantages of our baselines, which can be added to different foundation models. We also discuss the remaining challenges for low-resource vision and paths forward for future works.

2. Low-Resource Image Transfer Evaluation

To study low-resource vision tasks we cannot simply take a subset of existing data and pretend it is low-resource. This will not present the same challenges as are present in true low-resource data. Instead, we collect image data that is severely limited in its online availability. Specifically, we present our Low-Resource Image Transfer Evaluation (LITE) benchmark which considers three low-resource vision tasks. We examine the common challenges among these tasks and whether they can be solved by foundation models.

2.1. Tasks

Our benchmark has three tasks: (i) circuit diagram classification, (ii) image-to-image retrieval with historic maps, and (iii) image-to-image retrieval with mechanical drawings. Examples from each task are shown in Figure 2.

Task I: Circuit Diagram Classification. The goal is to classify the images of circuit diagrams by their function, *e.g.*, audio amplifier and power supply. We collect circuit images and labels from books [19] and websites [1–3]. In total, we have 32 function classes which are equally represented in training. The challenge comes from small changes in circuit components dramatically changing the function. Since there are different layouts for the same function, it is also easy for models to overfit to specific layouts. We measure performance with Top-1 and Top-5 accuracy.

Task II: Historic Map Retrieval. The task is to retrieve the corresponding modern-day satellite image for each image of a historic city map. Data is acquired from Old Maps Online [6] and cropping the corresponding contemporary satellite image from Google Maps [4]. This task is challenging as many city layouts have changed considerably over time and the contours of walls and buildings in the historic map may no longer exist in the satellite image. Moreover, historic maps originating from different regions and eras have vastly different cartographic styles. Performance is measured using Recall@1, Recall@5, and mean rank.

Task III: Mechanical Drawing Retrieval. The goal is to retrieve the mechanical drawing matching the image of a 3D-rendered component and vice versa. We collect mechanical drawings and rendered images from TraceParts [7] and GrabCAD [5]. The difficulty comes from the large visual difference between image sets. Moreover, the mechanical drawings and rendered images use different viewpoints. We evaluate this task with Recall@1, Recall@5, and mean rank, each averaged across both retrieval directions.

We summarize our benchmark statistics in Table 1. Note that we have collected as much data as we can find freely available online for each task, yet, the amount of data is still incredibly small showing how low-resource these tasks are.

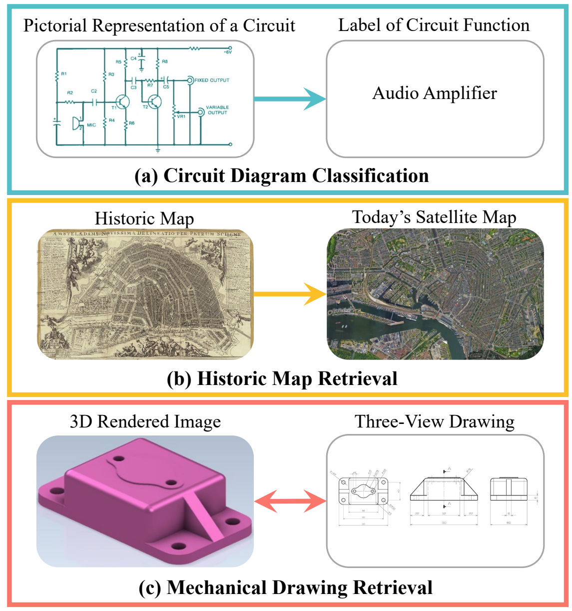

Figure 2. **Low-Resource Image Transfer Evaluation Benchmark**. Our three benchmark tasks are: (a) classifying circuit diagrams with the correct function, (b) retrieving the modern satellite map given an old map of a city, and (c) retrieving the mechanical drawing corresponding to a 3D photo of a component and vice versa.

Task	Formulation	Train	Val	Test
Circuit Diagram Classification	Image Classification	154	100	1,078
Historic Map Retrieval	Image-to-Image Retrieval	102	140	409
Mechanical Drawing Retrieval	Image-to-Image Retrieval	300	100	754

Table 1. **LITE Benchmark Statistics.** We show the task formulation and number of images (or image pairs) per split for each of our three tasks. The benchmark is available on our project website.

2.2. Low-Resource Vision Challenges

While the three low-resource tasks forming our LITE benchmark are very diverse, we identify three common challenges.

Challenge I: Data Scarcity. The data available for training models for low-resource scenarios is extremely limited. This is demonstrated through the small amount of data we were able to find online for each low-resource task (see Table 1).
Challenge II: Fine-Grained. Data that is low-resource is also highly specialized, meaning differences between images are incredibly subtle and attention to fine-grained details is necessary to solve the task. For example, the component symbols are key to a circuit's purpose, not its layout. Similarly, in mechanical drawings, the components may only vary in the number of holes.
Challenge III: Specialized Domain. Not only is the available data severely limited, but it has a significantly different appearance to the natural images commonly used in vision tasks. This means it is difficult to bootstrap the training data for low-resource tasks with existing datasets. Moreover, models that are successful on natural images cannot be easily applied to the specialized domains of low-resource images.

Each of these challenges has been studied in isolation in vision, for instance with few-shot learning [62, 72], fine-grained classification [29, 69, 88] and domain generalization [70]. However, their combination is unique to low-resource vision tasks. This means existing solutions to individual challenges cannot be easily applied to low-resource vision. Considering these challenges and their combination, we identify foundation models as the existing solution with the most potential to tackle low-resource vision, due to the impressive generalizability foundation models have shown. In the following section, we propose one way to better adapt foundation models for each low-resource vision challenge.

3. Baselines for the Low-Resource Challenges

Our goal is to adapt foundation models, pre-trained on large-scale datasets, to low-resource tasks. A foundation model $\mathcal{F}$ can be adapted with a small set of transfer learning parameters $\boldsymbol{\theta}$ as in LoRA [36] or AdaptFormer [16]. To better handle adaptation in low-resource vision, we introduce one baseline for each challenge highlighted in Section 2. First, to cope with the lack of data, we propose to augment training samples via generative models (Section 3.1). Second, to focus on the fine-grained details, we reduce the token patch size with selective tokenization (Section 3.2). Third, we introduce attention for specialized domains for better model adaption (Section 3.3). During finetuning on a low-resource task, we fix the foundation model $\mathcal{F}$ and train the parameters for transfer learning, our tokenization, and our attention.

3.1. Baseline I: Generated Data for Data Scarcity

Objective. Since a major challenge of low-resource vision is data scarcity, models are prone to overfitting the training data. We address this challenge by creating more training data for a low-resource vision task through generative models.
Novelty. Prior works have used generative models to produce realistic images to augment the training data [31, 65].

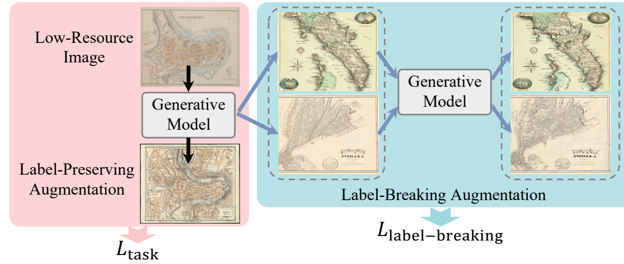

Figure 3. **Generated Data for Data Scarcity**. We augment images with generative models, obtaining images close to the input image where the label is preserved as well as more diverse images which break the label. We use label-preserving images in the task loss and augment the label-breaking images for use in a contrastive loss.

However, these works focus exclusively on data where the label of the augmented image is known. With this approach, it is challenging to achieve good data diversity with the highly limited number of images available in low-resource vision tasks. Therefore, besides label-preserving images, we use images where the original label is broken and unknown.
Method. Our proposed baseline is shown in Figure 3. In Stable Diffusion [59], the forward process gradually adds Gaussian noise to an image with a variance schedule $\beta_1, ..., \beta_T$ (T=50). To obtain new images, we sample noisy images at different timestep t and start the reverse process. For label-preserving augmentations we want a small t so the generated image is close to the original and thus adopt γ=0.3 and t=$\gamma \cdot T$. For label-breaking augmentations we use τ=0.6 and t=$\tau \cdot T$. Then, we can obtain various augmented images $[\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_m]$. Since the ground truth for the label-breaking augmentation is unknown we instead utilize such data with a contrastive learning objective [17, 30]. To construct the positive pairs, we generate a second augmented image $\mathcal{I}'_j$ for each label-breaking augmentation $\mathcal{I}_j$ with sampling timestep t=$\gamma \cdot T$. The contrastive loss encourages the feature of $\mathcal{I}'_j$ to be close to that of the label-breaking image $\mathcal{I}_j$, but far away from other label-breaking augmentations. We pass the label-breaking augmentation pairs through the foundation model to obtain their features x_j=$\mathcal{F}(\mathcal{I}_j)$ and x'_j=$\mathcal{F}(\mathcal{I}'_j)$, and our objective becomes:

$$L_{\text{label-breaking}} = -\frac{1}{N}\sum_{j}^{N} \log \frac{\exp(\mathbf{x}_j'^{\mathrm{T}}\mathbf{x}_j/\sigma)}{\sum_{i=1}^{N}\exp(\mathbf{x}_j'^{\mathrm{T}}\mathbf{x}_i/\sigma)}, \quad (1)$$

where N is the number of label-breaking image pairs and σ is the temperature for logit scaling. Combining this with the original task loss L_{task} our overall learning objective is:

$$L = L_{\text{task}} + \lambda L_{\text{label-breaking}}, \quad (2)$$

where λ is a hyperparameter to balance the loss terms. For L_{task} we use the original images as well as the label-preserving augmentations in a softmax cross-entropy for classification and a contrastive loss for retrieval. During each training iteration, we randomly sample B images from

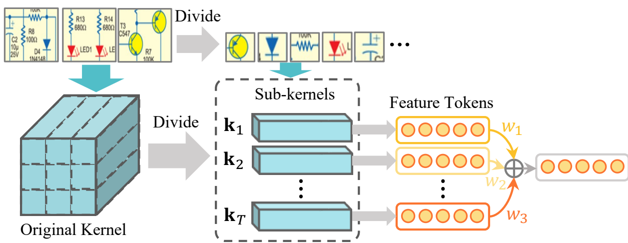

Figure 4. **Tokenization for Fine-Grained**. We divide the original linear projection of a pre-trained foundation model into sub-kernels. These sub-kernels can be applied to smaller areas of the image patch to attend to fine-grained details. We learn a weighting to combine the resulting features into patch-level features.

the union set of original and label-preserving augmentations for the task loss in addition to B image pairs for the label-breaking loss. Following [30] we use a memory bank for the contrastive loss $L_{\text{label-breaking}}$ so that there are N pairs ($N>B$) in total. As both the label-preserving and the label-breaking augmentations enlarge the data space for model learning, the challenge of limited data can be alleviated considerably.

Details. We use a batch size of B=8 and generate m=10 augmented images per sample. The contrastive learning uses a memory bank of size N=100 and balances loss terms with λ=0.1 (Eq. 2). We optimize using Adam [43] with a learning rate of 10^{-3} for 90 epochs on an A6000.

3.2. Baseline II: Tokenization for Fine-Grained

Objective. The second major challenge in low-resource vision, is the subtle, fine-grained details that distinguish different images. To address this challenge, we simply reduce the image patch size for tokenization so that the model can attend to the finer details of a low-resource input image.

Novelty. As we have limited data, we cannot train a new tokenization layer from scratch to reduce patch size. As shown in Figure 4, we instead divide the original linear projection kernel into sub-kernels which can be applied to smaller image patches. We then create patch-level features with a learned weighting. This allows attention to be paid to small local regions, crucial for fine-grained recognition [20, 21, 86], while only adding a handful of parameters.

Method. Vision foundation models [27, 44, 49, 58] divide the input image into large patches, *e.g.*, 16×16 or 14×14, so that the number of resulting tokens is small allowing training with large batch sizes. These image patches are linearly projected into features. The mechanism for this linear projection can be viewed as a convolution kernel $\mathbf{K} \in \mathbb{R}^{q\times q\times 3\times d_{\text{model}}}$ where $q\times q\times 3$ is also the dimensionality of an input image patch. We divide this kernel $\mathbf{K}$ into a series of sub-kernels $\{\mathbf{k}_1, \cdots, \mathbf{k}_T | \mathbf{k}_t \in \mathbb{R}^{u\times u\times 3\times d_{\text{model}}}\}$, where $u<q$. We use each sub-kernel to encode an input image patch and obtain a series of features, one per sub-kernel, $\{\mathbf{b}_1, \cdots, \mathbf{b}_T | \mathbf{b}_t \in \mathbb{R}^{p\times p\times d_{\text{model}}}\}$, where p is the size of feature maps. Unlike the original linear projection, the sub-kernels can find smaller, fine-grained patterns in the input image patch, achieving a similar effect to a reduced patch size. We learn a weighting $\mathbf{w}=[w_1, \cdots, w_T]$ to combine sub-kernel features into a patch-level feature $\mathbf{b}$:

$$\mathbf{b} = \sum_t \frac{e^{w_t}}{\sum_t e^{w_t}} \mathbf{b}_t. \tag{3}$$

To obtain output features with the same dimension as the original projection, we apply max pooling $\mathbf{b}'$=MaxPool($\mathbf{b}$) and flatten $\mathbf{b}'$. We can then use the existing positional encodings and class tokens and ensure our input is suitable for the frozen foundation model. As only the weighting $\mathbf{w} \in \mathbb{R}^T$ is learned, our tokenization is suitable for low-resource data and effectively encourages focus on fine-grained details.

Details. The initial kernel has size q=14. We set u=7, giving T=49 sub-kernels. The resulting feature maps of size p=32 are max pooled with a kernel size and stride of 2.

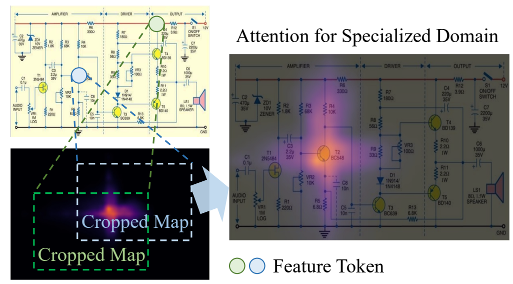

Figure 5. **Attention for Specialized Domains**. We learn a set of global attention maps with common attention patterns particular to the specialized domain such as vertical and horizontal directions for circuit diagrams. For each token, we crop the corresponding region from the global attention map according to the location.

3.3. Baseline III: Attention for Specialized Domains

Objective. The third challenge considers the adaption of foundation model features to the specialized low-resource domains. The transformer attention of foundation models struggles to distinguish the important regions in our specialized low-resource domains, we thus propose an alternative.

Novelty. We observe that the type of attention required is specific to each domain, but can be shared across different images and different patches within an image. For example, the vertical and the horizontal surroundings of a patch are important in circuit diagrams, while for historic maps, local neighbors are essential. To reduce the number of parameters, we share attention maps across samples and feature tokens by learning global attention maps. As shown in Figure 5, for each feature token, we simply crop the corresponding attention maps from the global maps.

Method. Specifically, we learn C attention maps $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{C\times 2h\times 2h}$, where h is the height and width of the feature maps before being flattened for input into the following transformer blocks. Each attention map will correspond to a different attention pattern. To obtain the correct size of $h\times h$ for a token's attention map we crop a sub-map from each global attention map. For a token corresponding to location (i, j), the top-left corner in the global attention map is $(h-i, h-j)$. As a result, we obtain one $h\times h$ sub-map for each of the h^2 tokens and form $\boldsymbol{\mathcal{M}}'_c \in \mathbb{R}^{h^2\times h\times h}$, for each of the C global attention maps. We flatten $\boldsymbol{\mathcal{M}}'_c$ into $\boldsymbol{\mathcal{M}}_c \in \mathbb{R}^{h^2\times h^2}$, and apply softmax to the last dimension. The resulting attention is multiplied with the values $\mathbf{V}$ used in the original multi-head self-attention. We weigh the resulting features with a learned vector $\mathbf{r}=[r_1, \cdots, r_C]$ as follows:

$$\hat{\mathbf{f}}_l = \sum_{c=0}^{C} \frac{e^{r_c}}{\sum_c e^{r_c}} \text{MLP}(\text{softmax}(\boldsymbol{\mathcal{M}}_c)\mathbf{V}), \tag{4}$$

where the multi-layer perceptron (MLP) is the same as used in the transformer layer's multi-head attention. We combine the output from our attention for specialized domains $\hat{\mathbf{f}}_l$ with the output from multi-head attention $\bar{\mathbf{f}}_l$ as:

$$\bar{\mathbf{f}}'_l = \bar{\mathbf{f}}_l + \alpha\hat{\mathbf{f}}_l, \tag{5}$$

where α is learned to balance the two attentions. As only $\boldsymbol{\mathcal{M}}$, α, and $\mathbf{r}$ are learned, training our attention for specialized domains allows adaptation without overfitting.

Details. We learn C=10 maps for the middle (16th) transformer block, leaving other blocks unchanged.

4. Related Work

High-Resource Vision. The large majority of computer vision research focuses on high-resource settings, where data is plentiful. Various benchmarks of high-resource images have been proposed, [18, 26, 47, 51, 54, 55, 60], unlocking the ability to train larger and larger models. Their images are crawled from the internet [47, 51, 54, 60], or captured by the authors [18, 26, 55]. The labels can be either coarse-grained [11, 18, 26, 51, 60, 80] or fine-grained [42, 47, 54, 55, 78]. High-resource vision tends to focus on natural images which are plentiful online. However, some benchmarks also collect images from other domains, *e.g*., X-ray [68], underwater [34], medical [38] and satellite [33]. These are less high-resource than natural images. However, they still contain thousands of samples. Different from previous high-resource image datasets, we focus on low-resource settings, where images are severely limited with only a few hundred samples available for training.

Vision Foundation Models. Vision foundation models are pre-trained by high-resource web-crawled images with weak supervision or human annotations, and present impressive generalizability on various downstream tasks. While CLIP [58], BLIP [49], and ALIGN [40] learn from image-text pairs only, ImageBind [27] uses image-paired data of multiple modalities. Recent works SAM [74], DINOv2 [56], UniDetector [44] and AIM [22] instead propose foundation models for visual-only tasks such as object detection, segmentation and depth estimation. However, the impressive generalization ability has been focused on natural images, likely similar to many in the large-scale training set [77]. Simultaneously, there are many low-resource problems from specialized domains lacking a large amount of online data, such as technical images. In this paper, we create a benchmark of low-resource vision problems and demonstrate that foundation models cannot generalize to such data.

It is also possible to adapt the strong image representations of foundational models to new tasks. This can be done by fintuning [79] or by training additional projection layers [25]. Several works [16, 28, 36, 52, 61] instead add new trainable parameters into the layers of a frozen pre-trained model. Although these works achieve impressive performance, they are not suited for low-resource vision where training data is severely limited and from fine-grained, specialized domains that are highly dissimilar to the pre-training data. We study such tasks and their challenges and propose baselines for better adaptation to low-resource tasks.

Low-Shot Vision. A huge number of works have studied scenarios with limited training data. One typical setting is few-shot learning [13, 48, 67], which aims to generalize to previously unseen classes with only a few training samples. Some works study in-context learning [10, 63, 64, 73, 84], which allows inference on unseen tasks by conditioning on related examples without updating the model parameters. Other works reduce this one step further and study zero-shot scenarios [9, 15, 45, 75, 76, 83], where no data of the relevant classes are seen in training, although prior knowledge or data from other classes could be used. All these works make a significant step towards reducing the amount of data needed for model learning. However, none of these tasks study the combination of scarce data, fine-grained differences and highly specialized domains present in low-resource vision.

5. Results and Discussion

5.1. Difficulties for Vision Foundation Models

To understand how well current vision foundation models address low-resource vision tasks, we first examine their zero-shot performance on our LITE benchmark. We consider six vision foundation models: CLIP [58], BLIP [49], SAM [44], AIM [22], DINOv2 [56] and ImageBind [27].

Setup. To obtain zero-shot results for circuit diagram classification, we follow [58] and customize the label text to make it better suited to the models. Specifically, we use the prompt template "A circuit diagram of {label}.", where the label is a category label, *e.g*., power supply or motor driver. We cannot obtain zero-shot results for SAM, AIM, and DINOv2 in this way, so we omit these models for circuit diagram classification. For all tasks, we calculate the similarities among the feature embeddings between the input image and the ground-truth image or text to find the closest neighbors.

Results. From Table 2 we observe none of these foundation models perform well on low-resource tasks. Although ImageBind is better suited due to its larger pre-training set and

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
CLIP [58]	7.7	**28.5**	**31.3**	**60.4**	**12.1**	3.6	10.5	210.2
BLIP [49]	**8.7**	28.2	2.2	12.5	52.1	2.5	7.8	209.4
SAM [44]	-	-	0.7	3.2	97.0	0.1	0.8	369.2
AIM [22]	-	-	12.0	33.0	37.9	**14.9**	**31.2**	72.2
DINOv2 [56]	-	-	1.5	7.1	83.4	**15.9**	**32.2**	**83.0**
ImageBind [27]	**19.3**	**45.1**	**28.1**	**62.1**	**10.1**	13.2	26.3	**83.1**

Table 2. **Difficulties for Vision Foundation Models**. We present zero-shot transfer performance. We mark the best in **red** and the second in **blue**. While ImageBind [27] has generally better zero-shot transfer ability on low-resource vision tasks, the tasks are far from solved.

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Simple Transformations		
Random Crop and Flip	19.8	45.3
Mixup [82]	20.8	46.0
CutMix [81]	20.0	45.5
Random Erasing [87]	20.8	46.2
Generative Models		
DA-Fusion [65]	19.6	45.1
SyntheticData [31]	20.8	46.0
Our Baselines		
Generated Data for Data Scarcity	**21.3**	**46.9**
Combination of Baselines	**24.1**	**49.3**

Table 3. **Challenge I: Data Scarcity**. We mark the best in **red** and the second in **blue**. Simple transformations do little to improve the diversity of training data. We obtain the best data diversity and thus the best baseline performance with our baselines which leverage both similar and dissimilar images produced by generative models.

image-focused embedding, there is still much room for improvement. Despite current foundation models' impressive generalizability on other benchmarks, they cannot yet solve the combined challenges of data scarcity, fine-grained details, and highly specialized domains. Unlike other zero-shot and few-shot tasks where foundation models have shown good generalization, low-resource data is truly scarce online, meaning it is unlikely to be in the training data of foundation models. The specialized domain means it is also highly dissimilar to natural images which form a large part of foundation model training data [77]. Due to the models' unfamiliarity with low-resource data, they struggle to attend to fine-grained, task-relevant details. Therefore, vision foundation models need adaptation for low-resource tasks.

5.2. Challenge Results

Setup. As ImageBind obtains the best zero-shot performance on our low-resource benchmark, we use it in this section. For all three challenges defined in Section 2 we add our proposed baselines or existing methods alongside AdaptFormer [16], keeping the foundation model frozen. Our baselines are independent of each other, focusing on different areas of the foundation model: input, tokenization, and attention. Thus they can be easily combined. We test this combination as well as the individual baselines.

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Fine-Grained		
Adaptive-FGSBIR [14]	16.7	43.2
PLEor [71]	17.1	44.1
PDiscoNet [66]	16.2	43.5
Our Baselines		
Tokenization for Fine-Grained	**20.9**	**45.5**
Combination of Baselines	**24.1**	**49.3**

Table 4. **Challenge II: Fine-Grained**. The **best** and **second** are highlighted. Fine-grained recognition methods need thousands of images for model learning, making them unsuited to low-resource tasks. Our tokenization baseline better utilizes the limited training data. However, there is much potential for further improvements.

Challenge I: Data Scarcity. Table 3, demonstrates the challenge of low-resource vision for existing solutions to data scarcity. Specifically, we test popular data augmentation methods on circuit classification. Traditional methods like random crop and flip as well as CutMix [81] struggle with our LITE benchmark. When using such a small set of images with very fine-grained differences these methods deliver limited additional data diversity. Mixup [82] and SyntheticData [31] obtain better performance as they can create more diverse training data by mixing samples and utilizing generative models. Although DA-Fusion [65] and SyntheticData [31] use generative models to obtain more data, they only consider generated images that are similar to the original ones, *i.e.* label-preserving. In contrast, our baseline considers both label-preserving and label-breaking generated images and therefore benefits from many more data points crucial for low-resource vision tasks. This is demonstrated further in the appendix with results for the other two tasks. Combining our baseline solutions to all three challenges results in the best performance, highlighting the multifaceted nature of low-resource vision.

Challenge II: Fine-Grained. We investigate how well recent state-of-the-art fine-grained methods [14, 66, 71] can tackle the challenge of low-resource vision in Table 4. We show results for circuit diagram classification here, results for the other two tasks can be found in the appendix. We use publicly available implementations except for PLEor [71],

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
CLIP [58]								
Zero-Shot Transfer	7.7	28.5	31.3	60.4	12.1	3.6	10.5	210.2
AdaptFormer	14.0	40.1	33.3	58.4	12.3	37.3	61.5	20.3
Our Baselines	**19.4**	**45.3**	**37.1**	**66.5**	**10.9**	**42.0**	**66.9**	**17.5**
BLIP [49]								
Zero-Shot Transfer	8.7	28.2	2.2	12.5	52.1	2.5	7.8	209.4
AdaptFormer	9.3	30.0	14.5	17.6	50.2	12.3	17.5	163.0
Our Baselines	**14.1**	**36.7**	**19.3**	**20.5**	**47.2**	**17.1**	**22.6**	**143.2**
SAM [44]								
Zero-Shot Transfer	-	-	0.7	3.2	97.0	0.1	0.8	369.2
AdaptFormer	18.0	44.6	8.6	12.8	59.3	7.4	10.2	221.3
Our Baselines	**22.4**	**47.8**	**13.4**	**16.3**	**55.4**	**11.2**	**16.0**	**178.7**
AIM [22]								
Zero-Shot Transfer	-	-	12.0	33.0	37.9	14.9	31.2	72.2
AdaptFormer	16.3	41.8	16.4	37.8	32.5	55.2	78.3	12.7
Our Baselines	**20.1**	**45.7**	**20.3**	**41.2**	**28.8**	**59.4**	**82.9**	**10.0**
DINOv2 [56]								
Zero-Shot Transfer	-	-	1.5	7.1	83.4	15.9	32.2	83.0
AdaptFormer	15.8	40.3	13.6	16.9	58.5	56.0	79.1	16.5
Our Baselines	**20.3**	**45.8**	**18.2**	**21.9**	**54.1**	**60.7**	**83.1**	**12.4**
ImageBind [27]								
Zero-Shot Transfer	19.3	45.1	28.1	62.1	10.1	13.2	26.3	83.1
AdaptFormer	19.8	45.5	30.3	62.6	13.4	54.3	76.6	13.8
Our Baselines	**24.1**	**49.3**	**36.4**	**68.0**	**9.8**	**60.0**	**82.5**	**10.2**

Table 5. **Combination with Different Foundation Models**. Our approach can easily be applied to different foundational models, improving their adaptation to low-resource tasks. However, the tasks are far from solved highlighting the need for further study of low-resource vision.

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Full-Parameter Finetuning	13.2	38.6
Transfer Learning		
Linear Probe	18.7	45.9
TOAST [61]	16.4	43.3
CLIP-Adapter [25]	16.3	42.9
IA3 [52]	18.2	45.4
VPT [41]	19.4	45.2
LoRA [36]	15.5	42.2
AdaptFormer [16]	19.8	45.5
Our Baselines		
Attention for Specialized Domains	**20.6**	**47.0**
Combination of Baselines	**24.1**	**49.3**

Table 6. **Challenge III: Specialized Domain**. **Red** marks the best and **blue** marks the second. State-of-the-art transfer learning methods focus on common natural images similar to the training data of foundation models, therefore they struggle with low-resource tasks. As a result, our simple baselines can easily lead to improvements.

which we re-implement ourselves. Since fine-grained methods assume there is sufficient data for model learning, they suffer from severe overfitting, degrading the performance of the zero-shot transfer. We are able to improve performance with our tokenization for fine-grained which attends to fine-grained differences with only a few additional parameters.

Challenge III: Specialized Domain. We consider several state-of-the-art transfer learning methods [16, 25, 36, 41, 52, 61] for adaptation to the specialized domains of our low-resource vision tasks. We show results in Table 6. All existing baselines struggle to improve over zero-shot transfer with only AdaptFormer giving a slight improvement. While more suited to limited data than the fine-grained methods current transfer learning approaches still struggle with the severely limited data of low-resource tasks. They are also not designed to attend to fine-grained details. Our attention for specialized domains enables better generalization while introducing minimal parameters. Combining all our low-resource baselines to consider all three major challenges further improves the result. However, this is only an initial step towards solving low-resource vision.

5.3. Our Baselines on Different Foundation Models

In Table 5, we demonstrate that our low-resource baselines can be plugged into different foundational models by adding them to CLIP [58], BLIP [49], SAM [44], AIM [22], DINOv2 [56] and ImageBind [27]. All six foundation models can be improved by a large margin with adaptation to low-resource tasks. For example, by adding AdaptFormer, we observe +33.7% R@1 for CLIP, +40.3% R@1 for AIM, +40.1% R@1 for DINOv2 and +41.1% R@1 for ImageBind on mechanical drawing retrieval. The adaptation allows the foundation model features to be better suited to the specific domain of a low-resource task and can thus distinguish images with distinctive patterns. Adding our simple baselines results in further improvements. For example, there is an additional +4.7% R@1 improvement for CLIP and DINOv2 and +5.7% for ImageBind on mechanical drawing retrieval.

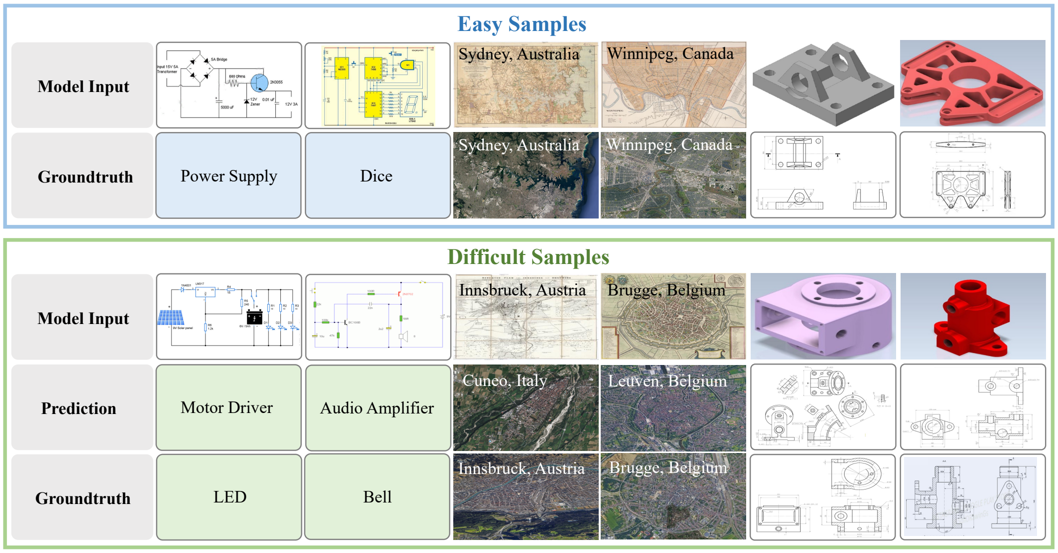

Figure 6. **Qualitative Results**. We show easy and difficult samples for our baselines. Our baselines can recognize prominent patterns in low-resource data, such as the coastline in the map of Sydney. However, they are overconfident, often basing predictions on one key region such as the presence of the battery in the LED circuit. Our baselines also cannot generalize to rarer image styles such as the Innsbruck map.

Nevertheless, the performance of our baselines with the best-performing foundation model is still low, *e.g.*, 24.1% Top-1 accuracy on circuit diagram classification and 37.1 R@1 on historic map retrieval. Thus, the proposed low-resource tasks are still far from solved and warrant further study.

5.4. Discussion

Qualitative Results. We present qualitative results in Figure 6. Our model successfully handles cases where a portion of the image is a clear indication of the label. For example, dice circuits contain a digital number display. For historic map retrieval, correct examples have a unique coastline or river path, while in mechanical drawings the correct component is clear from all views in the drawing. However, our baselines suffer when the relationships between multiple image regions are key. For instance, the horn in the bell circuit diagram also appears in audio amplifiers. The mechanical drawing failure cases appear correct from one drawing perspective but not the others. Our baselines also struggle when the image style is rare in training, as in the Innsbruck map.

Opportunities for Future Work. While our baselines have made a step towards adapting foundation models to low-resource vision tasks, these tasks are still far from solved. Our baselines still struggle to focus on informative regions due to the unfamiliar specialized domains, the fine-grained details within images, and the limited data we have to adapt foundation models. To better tackle the limited data, future works could focus on creating a greater diversity of generated data and explore whether seemingly irrelevant existing data could have some benefit to low-resource tasks. It is also important to consider the relationships between multiple image regions in order to make better fine-grained distinctions. To improve adaptation to specialized domains one possibility is to make the input data more suitable for foundation models with prompt learning or other techniques. Alternatively, future works could consider how foundation models can learn representations that are generalizable to non-natural images. In addition to these possible directions, there are also further challenges of low-resource vision beyond the three main challenges this paper explores. For instance, we consider the shift to specialized domains but not the domain shift between the input and ground truth or the sub-domains within the set of images. We also do not explicitly tackle the challenges of huge intra-class variation and imbalanced representation in the limited training set. Thus, there is still much room for further work on our low-resource benchmark.

6. Conclusion

This paper studies low-resource vision. We collect a benchmark of truly low-resource vision tasks and find these tasks share three challenges: extremely limited data, fine-grained differences between images, and highly specialized domains. To combat these challenges we investigate the generalization capability of foundation models, but find they struggle on low-resource vision tasks. We thus propose three baselines, one per challenge, in a step to solving low-resource vision. These baselines improve over prior works tackling individual challenges and can be easily plugged into different foundation models. Nevertheless, low-resource vision is still under-explored with many opportunities for future work.

Acknowledgement. This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

References

[1] Gadgetronicx. https://www.gadgetronicx.com/electronic-circuits-library/. 2, 16

[2] Circuit digest. https://circuitdigest.com/, . 16

[3] Circuits diy. https://www.circuits-diy.com/, . 2

[4] Google map. https://www.google.com/maps. 2, 16

[5] Grabcad. https://grabcad.com/. 2, 17

[6] Old maps online. https://www.oldmapsonline.org/. 2, 16

[7] Trace parts. https://www.traceparts.com/en. 2, 17

[8] Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. Cross-lingual word embeddings for low-resource language modeling". In *EACL*, 2017. 1

[9] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 38(7):1425–1438, 2015. 5

[10] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *NeurIPS*, 2022. 5

[11] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *ICCV*, 2023. 5

[12] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22): 2199–2210, 2017. 1

[13] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019. 5

[14] Ayan Kumar Bhunia, Aneeshan Sain, Parth Hiren Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive fine-grained sketch-based image retrieval. In *ECCV*, 2022. 6, 13

[15] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 5

[16] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition. In *NeurIPS*, 2022. 3, 5, 6, 7, 12, 13, 14

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[19] Suman Debnath. *270 Mini electronics project with circuit diagram*. 2015. 2, 16

[20] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019. 4

[21] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Zhanyu Ma, Yi-Zhe Song, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *ECCV*, 2020. 4

[22] Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Angel Bautista, Alexander Toshev, Vaishaal Shankar, Joshua M Susskind, and Armand Joulin. Scalable pretraining of large autoregressive image models. *arXiv preprint arXiv:2401.08541*, 2024. 5, 6, 7

[23] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *ACL*, 2017. 1

[24] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *TPAMI*, 28(4):594–611, 2006. 1

[25] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *IJCV*, pages 1–15, 2023. 5, 7, 13

[26] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5

[27] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 1, 4, 5, 6, 7

[28] Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-aware visual parameter-efficient fine-tuning. In *ICCV*, 2023. 5

[29] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *AAAI*, 2022. 3

[30] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 4

[31] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 3, 6, 13

[32] Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strotgen, and Dietrich Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *NAACL*, 2021. 1

[33] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 5

[34] Jungseok Hong, Michael Fulton, and Junaed Sattar. Trashcan: A semantically-segmented dataset towards visual detection of marine debris. *arXiv preprint arXiv:2007.08097*, 2020. 1, 5

[35] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *TPAMI*, 44(9):5149–5169, 2021. 1

[36] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3, 5, 7, 12, 13, 14

[37] Zixuan Hu, Li Shen, Zhenyi Wang, Tongliang Liu, Chun Yuan, and Dacheng Tao. Architecture, dataset and model-scale agnostic data-free meta-learning. In *CVPR*, 2023. 1

[38] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019. 5

[39] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[40] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 5

[41] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, 2022. 7, 13

[42] Faizan Farooq Khan, Xiang Li, Andrew J Temple, and Mohamed Elhoseiny. Fishnet: A large-scale dataset and benchmark for fish recognition, detection, and functional trait prediction. In *ICCV*, 2023. 5

[43] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4

[44] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1, 4, 5, 6, 7

[45] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017. 5

[46] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *ICML*, 2021. 1

[47] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5

[48] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 5

[49] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1, 4, 5, 6, 7

[50] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[51] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 5

[52] Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *NeurIPS*, 2022. 5, 7, 13

[53] Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *CVPR*, 2023. 1

[54] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[55] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. 5

[56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5, 6, 7

[57] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In *ACL*, 2017. 1

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 4, 5, 6, 7

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[60] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 5

[61] Baifeng Shi, Siyu Gai, Trevor Darrell, and Xin Wang. Toast: Transfer learning via attention steering. *arXiv preprint arXiv:2305.15542*, 2023. 5, 7, 13

[62] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 2023. 3

[63] Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*, 2023. 5

[64] Yasheng Sun, Yifan Yang, Houwen Peng, Yifei Shen, Yuqing Yang, Han Hu, Lili Qiu, and Hideki Koike. Imagebrush: Learning visual in-context instructions for exemplar-based image manipulation. *arXiv preprint arXiv:2308.00906*, 2023. 5

[65] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 3, 6, 13

[66] Robert van der Klis, Stephan Alaniz, Massimiliano Mancini, Cassio F Dantas, Dino Ienco, Zeynep Akata, and Diego Marcos. Pdisconet: Semantically consistent part discovery for fine-grained recognition. In *ICCV*, 2023. 6, 13

[67] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016. 1, 5

[68] Boying Wang, Libo Zhang, Longyin Wen, Xianglong Liu, and Yanjun Wu. Towards real-world prohibited item detection: A large-scale x-ray benchmark. In *ICCV*, 2021. 1, 5

[69] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *CVPR*, 2021. 3

[70] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 3

[71] Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, and Qi Tian. Open-set fine-grained retrieval via prompting vision-language evaluator. In *CVPR*, 2023. 6, 13

[72] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM Computing Surveys*, 53(3):1–34, 2020. 3

[73] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. *arXiv preprint arXiv:2305.01115*, 2023. 5

[74] Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, and Shengjin Wang. Detecting everything in the open world: Towards universal object detection. In *CVPR*, 2023. 5

[75] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 5

[76] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 5

[77] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 5, 6

[78] Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. Emoset: A large-scale visual emotion dataset with rich attributes. In *ICCV*, 2023. 5

[79] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 5

[80] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovskỳ, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araujo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *ICCV*, 2023. 5

[81] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 6, 13

[82] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 6, 13

[83] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 5

[84] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670*, 2023. 5

[85] Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Where is my spot? few-shot image generation via latent subspace optimization. In *CVPR*, 2023. 1

[86] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017. 4

[87] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 6, 13

[88] Haowei Zhu, Wenjing Ke, Dong Li, Ji Liu, Lu Tian, and Yi Shan. Dual cross-attention learning for fine-grained visual categorization and object re-identification. In *CVPR*, 2022. 3

[89] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, 2016. 1