

Low-Resource Vision Challenges for Foundation Models

Yunhua Zhang, Hazel Doughty, Cees G.M. Snoek
University of Amsterdam

Abstract

Low-resource settings are well-established in natural language processing, where many languages lack sufficient data for machine learning at scale. However, low-resource problems are under-explored in computer vision. In this paper, we strive to address this gap and explore the challenges of low-resource image tasks with vision foundation models. Thus, 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 the three challenges of 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 propose to 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 the three low-resource data sources in our benchmark demonstrate our proposals already provide a better baseline than common 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 website: <https://xiaobai1217.github.io/Low-Resource-Vision/>.

1. Introduction

Many have studied low-resource natural language processing [7, 21, 30, 54, 86], 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 [33, 35], few-shot learning [51, 64], or generative modeling [22, 82]. 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).

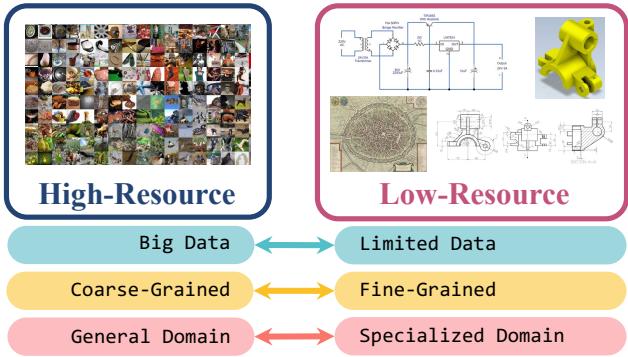


Figure 1. **High-Resource vs Low-Resource Vision.** High-resource vision focuses on images that can be collected at scale, have coarse-grained classes, and come from a general domain. We instead focus on low-resource vision tasks where data is scarce, has fine-grained differences, and comes from highly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

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 [32, 65] or even pairs [11, 44], the combination of all three is unique to low-resource vision and demands solutions outside of the scope of current models.

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 [25, 37, 42, 47, 55]. 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 [25, 47, 48, 55] 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 tasks are the lack of training data, fine-grained differences, and the domain shift from natural images to specialized domains. From our analysis, we find foundation models struggle to recognize and retrieve low-resource images as do existing transfer learning methods. Thus, we additionally introduce three simple baselines to mitigate each difficulty. Specifically, we finetune foundation models using diverse data produced from generative models to cope with the 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 recognition 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 the path 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, power supply, and signal generator. We manually collect the circuit images and labels from books [18] and websites [1, 2]. In total, we have 32 function classes which we ensure are equally represented in the training set. The challenge of this task comes from small changes in the circuit component giving the circuit a vastly different function. Since there are different layouts for the same function, it is also easy for models to overfit to specific layouts. To measure performance we report 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

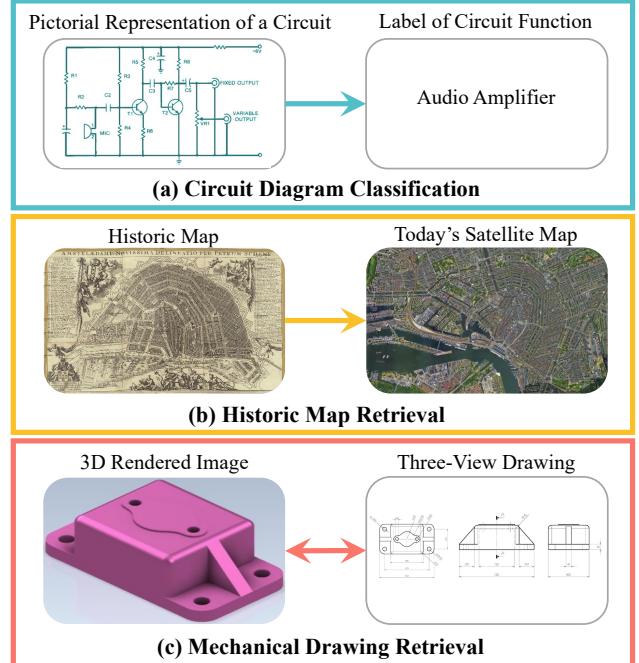


Figure 2. Low-Resource Image Transfer Evaluation Benchmark. Our three benchmark tasks are: (a) classifying circuit diagrams with the correct function, (b) retrieving the modern satellite map given an old map of a city, and (c) retrieving the mechanical drawing corresponding to a 3D photo of a component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

of a historic city map. Data is acquired from Old Maps Online [5] and cropping the corresponding contemporary satellite image from Google Maps [3]. 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 [6] and GrabCAD [4]. 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.

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

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 will be made publicly available.

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 and comes from an entirely different domain to the natural images commonly used in vision tasks. This means it is both difficult to bootstrap the training data for low-resource tasks with existing datasets and 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 [59, 69], fine-grained classification [27, 66, 85] and domain generalization [67]. 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 θ as in LoRA [34] or AdaptFormer [15]. 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.

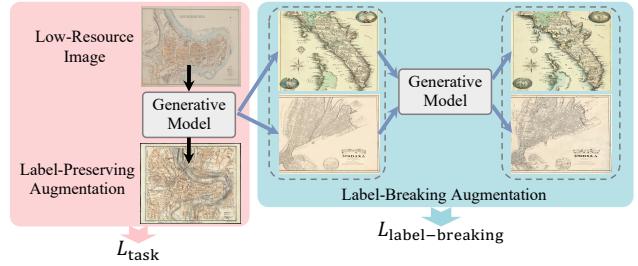


Figure 3. **Generated Data for Data Scarcity.** We augment images with generative models, obtaining images close to the input image where the label is preserved as well as more diverse images which break the label. We use label-preserving images in the task loss and augment the label-breaking images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

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 [29, 62]. 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. To obtain new images we apply different amounts of Gaussian noise onto a training image \mathcal{I} , before applying Stable Diffusion [56] to obtain various augmented images $[\mathcal{I}_1, \mathcal{I}_2, \dots, \mathcal{I}_m]$. We denote the amount of noise used to generate \mathcal{I}_j as n_j . With low noise, $n_j < \gamma$, the generated images are close to the original ones meaning the original label is preserved. With high noise, $n_j > \tau$, the augmentations become more extreme, often breaking the original label. Since the ground truth for the label-breaking augmentation is unknown we instead propose to utilize such data with self-supervised learning. For this, we adopt the contrastive learning objective [16, 28]. To construct the positive pairs, we generate a second augmented image \mathcal{I}'_j for each label-breaking augmentation \mathcal{I}_j with noise threshold $n_j < \gamma$. We use these images as a positive pair in a contrastive loss so that the feature of \mathcal{I}'_j is 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:

$$L_{\text{label-breaking}} = -\frac{1}{N} \sum_j^N \log \frac{\exp(\mathbf{x}'_j^\top \mathbf{x}_j / \sigma)}{\sum_{i=1}^N \exp(\mathbf{x}'_j^\top \mathbf{x}_i / \sigma)}, \quad (1)$$

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 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 [28] 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 $\gamma=0.3$ as the noise threshold for producing label-preserving augmentations and $\tau=0.6$ for label-breaking augmentations. We use a batch size of $B=8$ and generate $m=10$ augmented images per sample. The contrastive learning uses $N=100$ as the memory bank size and loss terms are balanced with $\lambda=0.1$ (Eq. 2). We optimize the losses using Adam [41] with a learning rating of 10^{-3} . The model is trained on one NVIDIA A6000 for 90 epochs.

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 [19, 20, 83], while only adding a handful of parameters.

Method. Vision foundation models [25, 42, 47, 55] 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 $\{\mathbf{k}_1, \dots, \mathbf{k}_T | \mathbf{k}_t \in \mathbb{R}^{u \times u \times 3 \times d_{\text{model}}}\}$, where $u < q$. We use each sub-kernel to encode an input image patch and obtain a series of features, one per sub-kernel, $\{\mathbf{b}_1, \dots, \mathbf{b}_T | \mathbf{b}_t \in \mathbb{R}^{p \times p \times d_{\text{model}}}\}$, 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

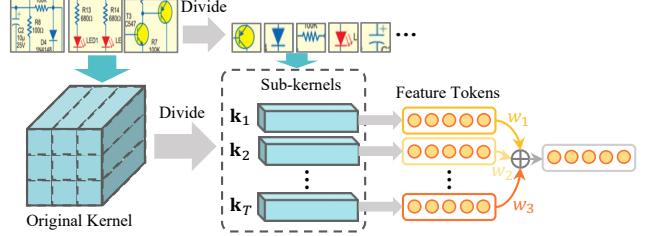


Figure 4. Tokenization for Fine-Grained. We divide the original linear projection of a pre-trained foundation model into sub-kernels. These sub-kernels can be applied to smaller areas of the image patch to attend to fine-grained details. We learn a weighting to combine the resulting features into patch-level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

patch size. We learn a weighting $\mathbf{w}=[w_1, \dots, w_T]$ to combine sub-kernel features into a patch-level feature \mathbf{b} :

$$\mathbf{b} = \sum_t \frac{e^{w_t}}{\sum_t e^{w_t}} \mathbf{b}_t. \quad (3)$$

To obtain output features with the same dimension as in the original projection, we apply max pooling $\mathbf{b}'=\text{MaxPool}(\mathbf{b})$ and flatten \mathbf{b}' . This allows us to use the existing positional encodings and class tokens and ensures our input is suitable for the frozen layers of the foundation model. Since only the weighting $\mathbf{w} \in \mathbb{R}^T$ is learned, our tokenization can be easily trained on low-resource data and effectively encourages the foundation model to 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.

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 $\mathcal{M} \in \mathbb{R}^{C \times 2h \times 2h}$, 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

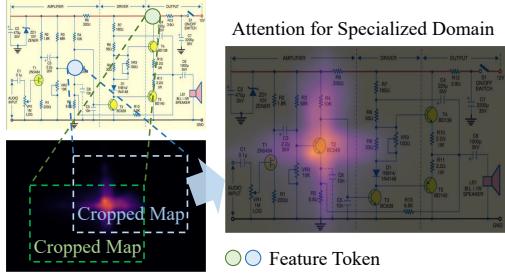


Figure 5. Attention for Specialized Domains. We learn a set of global attention maps with common attention patterns particular to the specialized domain such as vertical and horizontal directions for circuit diagrams. For each token, we crop the corresponding region from the global attention map according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

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 $\mathcal{M}'_c \in \mathbb{R}^{h^2 \times h \times h}$, for each of the C global attention maps. We flatten \mathcal{M}'_c into $\mathcal{M}_c \in \mathbb{R}^{h^2 \times h^2}$, 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 $\mathbf{r} = [r_1, \dots, r_C]$ as follows:

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

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:

$$\mathbf{f}'_l = \bar{\mathbf{f}}_l + \alpha \hat{\mathbf{f}}_l, \quad (5)$$

where α is a learnable parameter to balance the two attentions. As the only learnable parameters are \mathcal{M} , α , and \mathbf{r} , 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, *e.g.*, [17, 24, 45, 49, 52, 53, 57], unlocking the ability to train larger and larger models. Their images are crawled from the internet [45, 49, 52, 57], or captured by the authors [17, 24, 53]. The labels can be either coarse-grained [10, 17, 24, 49, 57, 77] or fine-grained [40, 45, 52, 53, 75]. High-resource vision tends to focus mostly on natural images which are the most plentiful online. However, some benchmarks also collect images from other domains, *e.g.*, X-ray [65], underwater [32], medical [36] and satellite [31]. These are less high-resource than natural images. However, they still contain at least several thousands of samples per benchmark. Different from

previous high-resource image datasets, we focus on the 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 (like captions) or human annotations, and present impressive generalizability on various downstream tasks. While CLIP [55], BLIP [47] and ALIGN [38] learn from image-text pairs only, ImageBind [25] uses image-paired data of multiple modalities. The recent SAM [71] and UniDetector [42] instead propose foundation models for object detection and segmentation. However, the impressive generalization ability has been focused on natural images, likely similar to many in the large-scale training set [74]. Simultaneously, there are many low-resource problems from specialized image 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 representation learned by foundational models to new tasks. This can be simply done by model fintuning [76] or by training an additional projection layer [23] on the pre-trained model features. Several works [15, 26, 34, 50, 58] instead add new parameters into the layers of a pre-trained model, only training these parameters. Although these works achieve impressive performance, they are not suited for low-resource vision where the training data is severely limited and from fine-grained, specialized domains that are highly dissimilar to the pre-training data. We propose to 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 [12, 46, 64], where the goal is to generalize to previously unseen classes with only a few samples available for model learning. Some works study the related task of in-context learning [9, 60, 61, 70, 81], 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 [8, 14, 43, 72, 73, 80], 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

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
CLIP [55]	7.7	28.5	31.3	60.4	12.1	3.6	10.5	210.2
BLIP [47]	8.7	28.2	2.2	12.5	52.1	2.5	7.8	209.4
SAM [42]	-	-	0.7	3.2	97.0	0.1	0.8	369.2
ImageBind [25]	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 [25] has generally better zero-shot transfer ability on low-resource vision tasks, the tasks are far from solved.

zero-shot performance on our LITE benchmark. We consider four vision foundation models that are publicly available: CLIP [55], BLIP [47], SAM [42] and ImageBind [25].

Setup. To obtain zero-shot results for circuit diagram classification, we follow [55] 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 [42] in this way, so we omit it for circuit diagram classification. For all tasks, we simply 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 the results in Table 2 we observe none of these foundation models perform well on low-resource settings. Although ImageBind is better suited due to its larger pre-training set size and image-focused embedding, there is still much room for improvement in low-resource vision. 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. This means it is unlikely to be present 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 [74]. Due to the model’s unfamiliarity with low-resource data, it struggles 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 transfer performance on our low-resource benchmark, we adapt this model by the existing transfer learning method AdaptFormer [15] in the following experiments. For all three challenges defined in Section 2 we add our proposed baselines or existing methods to this default setting, keeping the foundation model parameters frozen. Our baselines work independently 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.

Challenge I: Data Scarcity. In Table 3, we demonstrate

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Simple Transformations		
Random Crop and Flip	19.8	45.3
Mixup [79]	20.8	46.0
CutMix [78]	20.0	45.5
Random Erasing [84]	20.8	46.2
Generative Models		
DA-Fusion [62]	19.6	45.1
SyntheticData [29]	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.

the challenge of low-resource vision for existing solutions to data scarcity. Specifically, we test popular data augmentation methods on circuit diagram classification. Traditional data augmentation methods like random cropping and flipping as well as CutMix [78] struggle with our low-resource benchmark. When using such a small set of images with very fine-grained differences these methods deliver limited improvements to the data diversity. Mixup [79] and SyntheticData [29] obtain better performance as they are able to create more diverse training data by mixing samples and utilizing generative models. Although DA-Fusion [62] and SyntheticData [29] use generative models to obtain more data points, 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 can therefore benefit from many more data points crucial for low-resource vision tasks. This is demonstrated further in the appendix with the 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 [13, 63, 68] can tackle the challenge of low-resource vision in Table 5. We show results for circuit diagram classification here, results for the other two tasks can be found in the appendix. We use

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
CLIP [55]								
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 [47]								
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 [42]								
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
ImageBind [25]								
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 4. **Combination with Different Foundation Models.** Our approach can easily be plugged into different foundational models, improving their ability to adapt to low-resource tasks. However, the tasks are still far from solved demonstrating the need for further study of low-resource vision.

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Fine-Grained		
Adaptive-FGSBIR [13]	16.7	43.2
PLEor [68]	17.1	44.1
PDiscoNet [63]	16.2	43.5
Our Baselines		
Tokenization for Fine-Grained	20.9	45.5
Combination of Baselines	24.1	49.3

Table 5. **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.

publicly available implementations except for PLEor [68], 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 [15, 23, 34, 39, 50, 58] 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-

	Circuit Classification	
	Top-1 (%) ↑	Top-5 (%) ↑
Zero-Shot Transfer	19.3	45.1
Transfer Learning		
Linear Probe	18.7	45.9
TOAST [58]	16.4	43.3
CLIP-Adapter [23]	16.3	42.9
IA3 [50]	18.2	45.4
VPT [39]	19.4	45.2
LoRA [34]	15.5	42.2
AdaptFormer [15]	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.

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 4, we demonstrate that our low-resource baselines can be plugged into different foundational models by adding them to CLIP [55], BLIP [47], SAM [42] and ImageBind [25]. We observe that all four 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 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

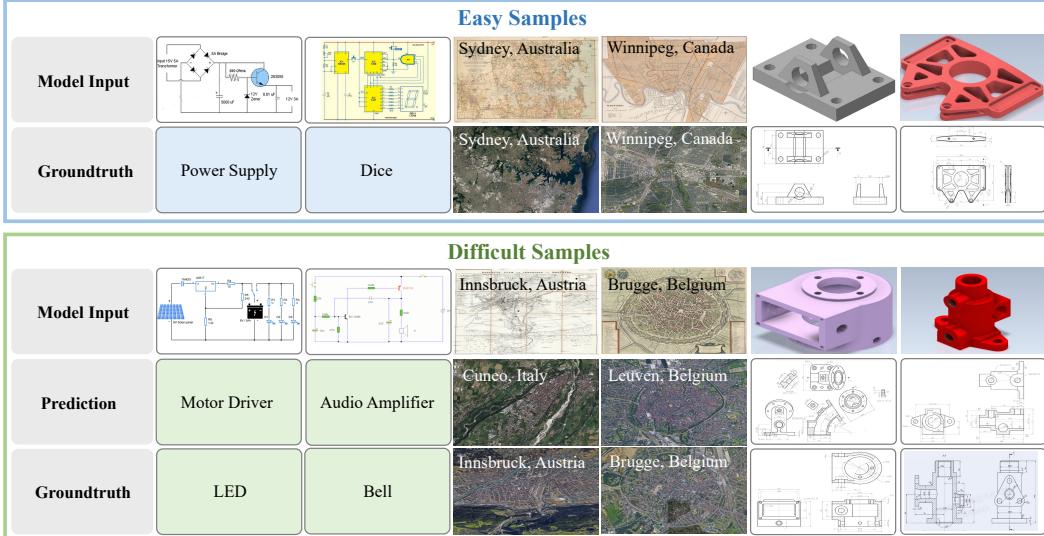


Figure 6. **Qualitative Results.** We show easy and difficult samples for our baselines. Our baselines can recognize prominent patterns in low-resource data, such as the coastline in the map of Sydney. However, they are overconfident, often basing predictions on one key region such as the presence of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

results in further improvements. For example, there is an additional +4.7% R@1 improvement for CLIP and +5.7% for ImageBind on mechanical drawing retrieval. 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 certain portion of the image is a clear indication of the groundtruth. For example, dice circuits usually 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 amplifier circuits. The mechanical drawing failure cases appear correct from one drawing perspective but not the others. Our baselines also struggle when image style is rare or unseen in training, as in the case of 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 tasks. We collect a benchmark of truly low-resource vision tasks and find these tasks share three common 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, to make a step toward solving low-resource vision. These baselines improve over prior works which tackle individual challenges in isolation and can be easily plugged into different foundation models. Nevertheless, this is still an under-explored area 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, 15
- [2] Circuit digest. <https://circuitdigest.com/>. 2, 15
- [3] Google map. <https://www.google.com/maps>. 2, 15
- [4] Grabcad. <https://grabcad.com/>. 2, 16
- [5] Old maps online. <https://www.oldmapsonline.org/>. 2, 15
- [6] Trace parts. <https://www.traceparts.com/en>. 2, 16
- [7] Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. Cross-lingual word embeddings for low-resource language modeling". In *EACL*, 2017. 1
- [8] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 38(7):1425–1438, 2015. 5
- [9] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. Visual prompting via image inpainting. In *NeurIPS*, 2022. 5
- [10] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinand, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *ICCV*, 2023. 5
- [11] 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
- [12] Luca Bertinetto, Joao F Henriques, Philip HS Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *ICLR*, 2019. 5
- [13] 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, 7, 13
- [14] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 5
- [15] 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
- [16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3
- [17] 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
- [18] Suman Debnath. *270 Mini electronics project with circuit diagram*. 2015. 2, 15
- [19] Yao Ding, Yanzhao Zhou, Yi Zhu, Qixiang Ye, and Jianbin Jiao. Selective sparse sampling for fine-grained image recognition. In *ICCV*, 2019. 4
- [20] 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
- [21] Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. In *ACL*, 2017. 1
- [22] Li Fei-Fei, Robert Fergus, and Pietro Perona. One-shot learning of object categories. *TPAMI*, 28(4):594–611, 2006. 1
- [23] 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
- [24] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 5
- [25] 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
- [26] Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-aware visual parameter-efficient fine-tuning. In *ICCV*, 2023. 5
- [27] 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
- [28] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3, 4
- [29] 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
- [30] 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
- [31] 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
- [32] 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
- [33] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *TPAMI*, 44(9):5149–5169, 2021. 1
- [34] 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

- [35] 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
- [36] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, 2019. 5
- [37] Chao Jia, Yafei 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
- [38] Chao Jia, Yafei 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
- [39] 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
- [40] 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
- [41] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 4
- [42] 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
- [43] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017. 5
- [44] 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
- [45] 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
- [46] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 5
- [47] 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
- [48] 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
- [49] 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
- [50] 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
- [51] 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
- [52] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5
- [53] 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
- [54] 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
- [55] 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
- [56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3
- [57] 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
- [58] Baifeng Shi, Siyu Gai, Trevor Darrell, and Xin Wang. Toast: Transfer learning via attention steering. *arXiv preprint arXiv:2305.15542*, 2023. 5, 7, 13
- [59] 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
- [60] 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
- [61] 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
- [62] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 3, 6, 13
- [63] 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, 7, 13
- [64] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NIPS*, 2016. 1, 5
- [65] 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

- [66] 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
- [67] 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
- [68] 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, 7, 13
- [69] 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
- [70] 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
- [71] 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
- [72] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning—the good, the bad and the ugly. In *CVPR*, 2017. 5
- [73] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 5
- [74] 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
- [75] 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
- [76] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NeurIPS*, 2014. 5
- [77] Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovský, Pelin Dogan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ondřej Chum, and André Araújo. Towards universal image embeddings: A large-scale dataset and challenge for generic image representations. In *ICCV*, 2023. 5
- [78] 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
- [79] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 6, 13
- [80] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 5
- [81] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. What makes good examples for visual in-context learning? *arXiv preprint arXiv:2301.13670*, 2023. 5
- [82] 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
- [83] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017. 4
- [84] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *AAAI*, 2020. 6, 13
- [85] 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
- [86] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer learning for low-resource neural machine translation. In *EMNLP*, 2016. 1

Low-Resource Vision Challenges for Foundation Models

Appendix

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
Zero-Shot Transfer	19.3	45.1	28.1	62.1	10.1	13.2	26.3	83.1
Linear Probe	18.7	45.9	-	-	-	-	-	-
LoRA [34]								
LoRA	15.5	42.2	34.0	69.2	9.1	41.8	66.7	19.0
+ Generated Data for Data Scarcity	18.1	44.6	35.7	71.0	8.7	43.2	68.9	17.2
+ Tokenization for Fine-Grained	19.8	46.0	36.7	72.2	8.6	44.9	70.2	16.1
+ Attention for Specialized Domains	21.0	47.3	37.9	73.3	8.4	46.4	72.3	14.8
AdaptFormer [15]								
AdaptFormer	19.8	45.5	30.3	62.6	13.4	54.3	76.6	13.8
+ Generated Data for Data Scarcity	21.3	47.0	33.7	64.7	11.5	57.2	79.1	12.0
+ Tokenization for Fine-Grained	22.7	48.1	34.9	66.4	10.8	58.8	81.2	11.4
+ Attention for Specialized Domains	24.1	49.3	36.4	68.0	9.8	60.0	82.5	10.2

Table 7. **Combination of Our Baselines.** Our generated data for data scarcity can mitigate the overfitting considerably for both LoRA and AdaptFormer. The attention for specialized domains and tokenization for fine-grained further contribute to performance improvement by helping the model focus on task-relevant regions and fine-grained details.

A. Combination of Our Baselines

Method. We adapt the foundation model ImageBind by an existing transfer learning method, and add our proposed baselines for the three challenges defined in Section 2. We keep the foundation model parameters frozen and only update the parameters introduced by the transfer learning method and our baselines. Our baselines work independently of each other, focusing on different areas of the foundation model: input, tokenization, and attention. Thus, they can be easily combined. Specifically, our generated data for data scarcity produce more samples for model learning, while our tokenization for fine-grained better encodes the input image patches into feature tokens. As the transfer learning method learns additional parameters inside the foundation model for the low-resource tasks, our attention for specialized domains is inserted into one layer to help the model focus on task-relevant regions.

Results. In Table 7, we ablate the effect of our three proposed baselines (in Section 3), *i.e.*, generated data for data scarcity, tokenization for fine-grained and attention for specialized domains. We consider both LoRA [34] and AdaptFormer [15] as the additional transfer learning parameters. By adding our generated data, the overfitting issue from limited training data can be alleviated considerably. For instance, this gives +2.6% Top-1 accuracy on circuit diagram classification with LoRA and 3.4% in R@1 on historic maps with AdaptFormer. This is because the label-breaking images enlarge the data space to help the representation learning. With our tokenization for fine-grained, the performance is further improved by +1.7% Top-1 with LoRA and +1.4% with AdaptFormer on circuit diagram classification. This

demonstrates the benefit of processing smaller regions compared to the original kernel so that the fine-grained details can be discovered. Adding attention for specialized domains to combat the out-of-distribution challenge delivers a further +1.2% Top-1 with LoRA and 1.4% with Adaptformer on circuit classification. Historic map retrieval and mechanical drawing retrieval obtain similar improvements. Our baselines are effective additions to both LoRA and AdaptFormer, increasing the results on circuit classification by 5.5% Top-1 and 4.3% respectively, with similar improvements for historic map retrieval and mechanical drawing retrieval.

B. Challenge Results on All Tasks

In Section 5.2 in the main paper, we demonstrate the challenges of low-resource vision for existing solutions on circuit diagram classification. Here, we present the same experiments on all three low-resource tasks including historic map and mechanical drawing retrieval.

Challenge I: Data Scarcity. In Table 8, we show the challenge of low-resource vision for existing solutions to data scarcity. Existing approaches are effective for mechanical drawing retrieval but give little improvement over zero-shot transfer on circuit diagram classification and historic map retrieval. Our generated data for data scarcity benefits all three tasks, giving the most improvement on historic map retrieval and mechanical drawing retrieval. This is because the domain gap between natural images and the data of these two tasks is slightly smaller making the label-breaking augmentations look more realistic (we show visualizations in Section E). Our generated data for data scarcity thus increases the data diversity more effectively for historic map and mechanical drawing retrieval. While our combination

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
Zero-Shot Transfer	19.3	45.1	28.1	62.1	10.1	13.2	26.3	83.1
Simple Transformations								
Random Cropping + Random Flipping	19.8	45.3	30.3	62.6	13.4	54.3	76.6	13.8
Mixup [79]	20.8	46.0	28.4	55.7	15.2	50.5	72.9	14.9
CutMix [78]	20.0	45.5	28.4	62.6	12.5	54.9	78.1	13.1
Random Erasing [84]	20.8	46.2	23.5	49.6	17.1	54.6	78.0	13.2
Generative Models								
DA-Fusion [62]	19.6	45.1	29.8	61.3	14.7	54.5	77.6	13.5
SyntheticData [29]	20.8	46.0	30.4	62.9	12.6	55.9	78.9	13.0
Our Baselines								
Generated Data for Data Scarcity	21.3	46.9	33.7	64.7	11.5	57.2	79.1	12.0
Combination of Baselines	24.1	49.3	36.4	68.0	9.8	60.0	82.5	10.2

Table 8. **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 on all the three low-resource tasks with our baselines which leverage both similar and dissimilar images produced by generative models.

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
Zero-Shot Transfer	19.3	45.1	28.1	62.1	10.1	13.2	26.3	83.1
Fine-Grained								
Adaptive-FGSBIR [13]	16.7	43.2	5.3	12.6	55.1	5.6	17.1	70.2
PLEor [68]	17.1	44.1	4.6	11.8	56.5	5.0	16.9	71.3
PDiscoNet [63]	16.2	43.5	5.8	13.0	53.2	5.4	17.2	69.8
Our Baselines								
Tokenization for Fine-Grained	20.9	45.5	32.1	64.0	12.3	55.9	78.4	12.7
Combination of Baselines	24.1	49.3	36.4	68.0	9.8	60.0	82.5	10.2

Table 9. **Challenge II: Fine-Grained.** We mark the best in red and the second in blue. 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 and makes improvements on all the three low-resource tasks. However, there is much potential for further improvements.

	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval			GFLOPs	Params (M)
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓		
Zero-Shot Transfer	19.3	45.1	28.1	62.1	10.1	13.2	26.3	83.1	224.6	63.3
Transfer Learning										
TOAST [58]	16.4	43.3	4.2	11.5	52.4	4.4	16.3	69.0	476.2	73.8
CLIP-Adapter [23]	16.3	42.9	3.2	15.9	42.1	7.7	23.2	59.5	224.8	64.2
IA3 [50]	18.2	45.4	29.1	51.3	19.5	52.0	76.7	12.7	224.6	63.6
VPT [39]	19.4	45.2	36.2	61.6	13.3	47.7	72.4	13.5	233.3	63.8
LoRA w/ Our Baselines										
LoRA [34]	15.5	42.2	34.0	69.2	9.1	41.8	66.7	19.0	224.7	63.7
+ Attention for Specialized Domains	16.9	44.5	35.1	70.9	8.8	43.0	69.2	17.8	224.7	63.7
+ Combination of Baselines	21.0	47.3	37.9	73.3	8.4	46.4	72.3	14.8	233.7	63.7
AdaptFormer w/ Our Baselines										
AdaptFormer [15]	19.8	45.5	30.3	62.6	13.4	54.3	76.6	13.8	224.6	63.4
+ Attention for Specialized Domains	20.6	47.0	31.9	64.2	12.1	56.4	78.3	12.8	224.6	63.4
+ Combination of Baselines	24.1	49.3	36.4	68.0	9.8	60.0	82.5	10.2	233.6	63.4

Table 10. **Challenge III: Specialized Domain.** We mark the best in red and the second in blue. 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 on all the three low-resource tasks.

of baselines does deliver improvements on all three tasks, they are still far from solved, uncovering the difficulties of low-resource vision.

Challenge II: Fine-Grained. We investigate how well recent state-of-the-art fine-grained methods [13, 63, 68] can tackle the challenge of low-resource vision in Table 9. While existing fine-grained methods assume there is sufficient data for model learning, they suffer from severe overfitting. This is demonstrated best on historic map and mechanical draw-

ing retrieval where the results of existing fine-grained methods are much lower than zero-shot transfer. In contrast, our baseline for fine-grained attends to fine-grained differences with only a few additional parameters so that the performance can be improved on all three tasks.

Challenge III: Specialized Domain. We consider several state-of-the-art transfer learning methods [15, 23, 34, 39, 50, 58] for adaptation to the specialized domains of our low-resource vision tasks. We show results in Table 10. Since our

Label-Preserving	Label-Breaking	Circuit Diagram Classification		Historic Map Retrieval			Mechanical Drawing Retrieval		
		Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓
✓		19.8	45.5	30.3	62.6	13.4	54.3	76.6	13.8
	✓	20.8	46.1	32.1	63.8	12.1	56.4	78.0	12.8
	✓	20.4	46.0	32.4	63.5	12.5	55.7	77.5	13.0
✓	✓	21.3	46.9	33.7	64.7	11.5	57.2	79.1	12.0

Table 11. **Ablation of Generated Data.** Both label-preserving and label-breaking generated images add more data points into the training data for model learning. Thus, both types of augmentation contribute to the reduction in overfitting.

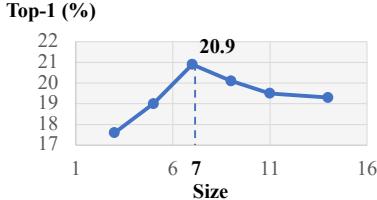


Figure 7. **Effect of Sub-Kernel Size** in tokenization for fine-grained. A medium sub-kernel gives a good trade-off between meaningful regions and fine-grained details.

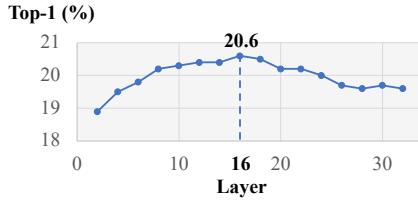


Figure 8. **Position of Attention** for specialized domains. The attention is reasonably robust to the choice of layer, although the middle layers reach a good trade-off between low- and high-level features.

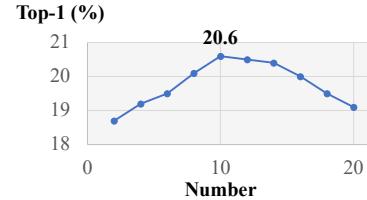


Figure 9. **Number of Attention Maps** for specialized domains. Learning anywhere from 8 to 14 attention maps allows the model to specialize to the domain while avoiding overfitting

baselines can be used in combination with any transfer learning method we plug them into two such methods: LoRA [34] and AdaptFormer [15]. While AdaptFormer performs better than other transfer learning methods on circuit diagram classification and mechanical drawing retrieval, LoRA is favorable on historic map retrieval. With our baselines, the results are improved further without much computation burden or parameters added. However, there is still no single model that always surpasses the others on all three low-resource tasks. Thus, this is only an initial step towards solving the challenges of low-resource vision.

C. Ablations and Hyperparameter Analysis

C.1. Generated Data for Data Scarcity

Effect of Label-Preserving and Label-Breaking Images. Our generated data for data scarcity in Section 3.1 use two types of augmentations for model learning, *i.e.*, label-preserving and label-breaking. While we adopt a supervised learning objective for label-preserving images with groundtruth labels, a self-supervised contrastive learning objective is applied to label-breaking images. In Table 11, we ablate their effect on our low-resource benchmark. Both types of images improve the diversity of training data. Thus, they reduce the overfitting and improve the model adaptation to our low-resource settings considerably. We provide visualizations of these two types of augmentations in Section E.

C.2. Tokenization for Fine-Grained

Effect of Sub-Kernel Size. In Section 3.2, we introduce our tokenization for fine-grained baseline, where the original kernel for linear projection is divided into sub-kernels for

encoding the input image patches. As the sub-kernels have smaller receptive fields, more simple patterns are encoded from the smaller input image patches. Here, we ablate the effect of sub-kernel size in Figure 7 on circuit diagram classification. While small sub-kernel sizes cannot cover meaningful regions and therefore only extract the texture information, large sub-kernel sizes focus more on global information than fine-grained details. Thus, adopting a medium size, *i.e.*, 7×7 , delivers the best performance.

C.3. Attention for Specialized Domains

In Section 3.3, we introduce our attention for specialized domains. Here, we discuss the effect of its position as well as the number of maps.

Position of Attention for Specialized Domains. We add our attention for specialized domains into only one layer to avoid introducing many additional parameters, which can result in overfitting. In Figure 8 we measure the effect of adding our attention to different transformer layers in circuit classification. As shallow layers focus on low-level, simple patterns, deep layers extract semantic features. The middle layers reach a trade-off between low- and high-level features. Nonetheless, our attention for specialized domains is reasonably robust to the choice of layer.

Number of Attention Maps. Using attention for specialized domains in the middle layer, we further study the effect of the number of attention maps C on circuit diagram classification in Figure 9. Using as few as two attention maps doesn't allow the model to fully specialize to the low-resource domain, while with 20 maps, the model overfits to the training data and cannot generalize to the variations seen at inference.

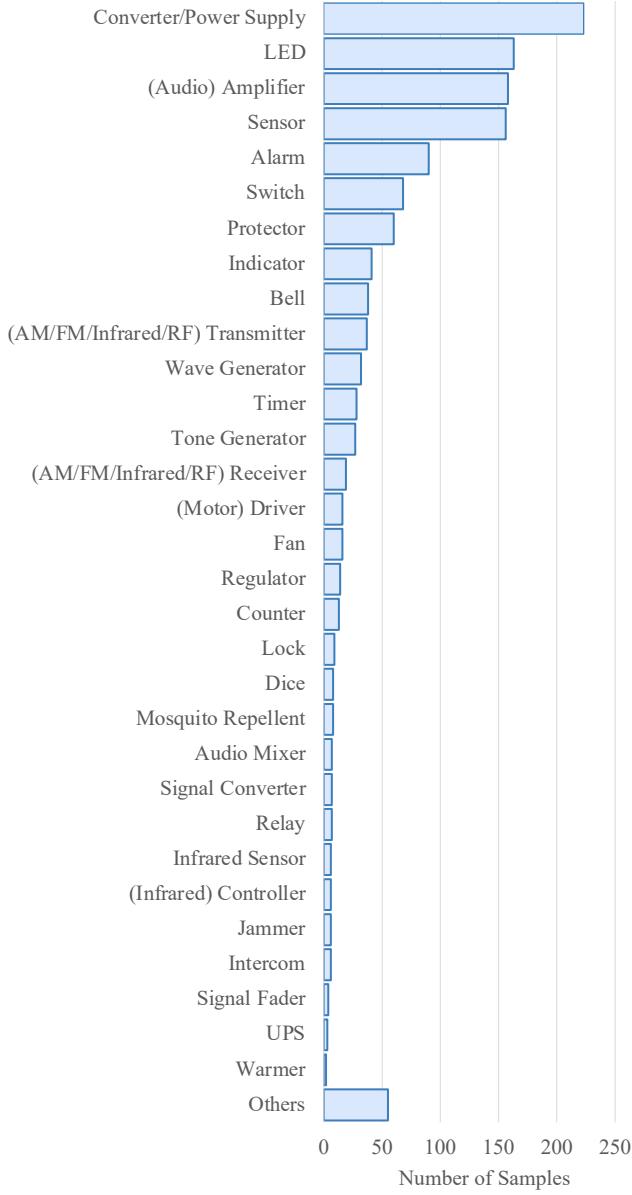


Figure 10. **Class Distribution of Circuit Diagrams.** While power supply, LED, amplifier and sensor have the most samples, it is hard to collect many circuit diagrams for most classes.

Learning 10 maps reaches the best trade-off, although the model is beneficial with any choice between 8 and 14.

D. Low-Resource Image Transfer Evaluation

In Section 2 in the main paper, we introduce our Low-Resource Image Transfer Evaluation benchmark. Here, we provide more details about our three low-resource tasks.

Task I: Circuit Diagram Classification. We collect 297 circuit diagrams from [18], 175 from Gadgetronix [1] and 860 from Circuit Digest [2]. This results in a total of 1,332 circuit diagrams. We divide them into 32 classes and present



Figure 11. **Country Distribution of Historic Maps.** While we find Europe and United States have the most historic maps online, other regions have much fewer maps available.

the class distribution in Figure 10. The power supply contains 223 samples, which is the most among all classes. We also find many samples for LED with 163 diagrams, amplifier with 158, and sensor with 156. However, we also find it hard to collect circuit diagrams for many classes, *e.g.*, relay, jammer, intercom, and signal fader. For instance, we can only get 7 samples of relays and 6 depicting jammers.

Task II: Historic Map Retrieval. All the historic maps come from OLD MAPS ONLINE [5] and all the satellite maps are from Google Map [3]. We collect 651 pairs of historic maps and today's satellite images from 29 countries and show the distribution across countries in Figure 11. United States and Europe have the most historic maps online. For example, we obtain 194 images for the United States. In contrast, collecting maps from other regions is more difficult. For instance, we can only find 2 images for Australia and 4

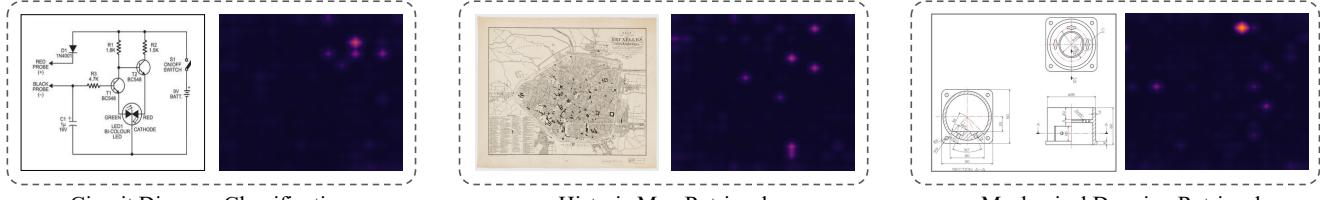


Figure 12. Attention Maps from Vision Foundation Model. We show the highest activation on each region across all the attention maps of the middle transformer block. Only a few regions are activated for the three images. This means that the vision foundation model fails to understand the interaction between different image regions. Thus, vision foundation models need proper adaptation for low-resource tasks.

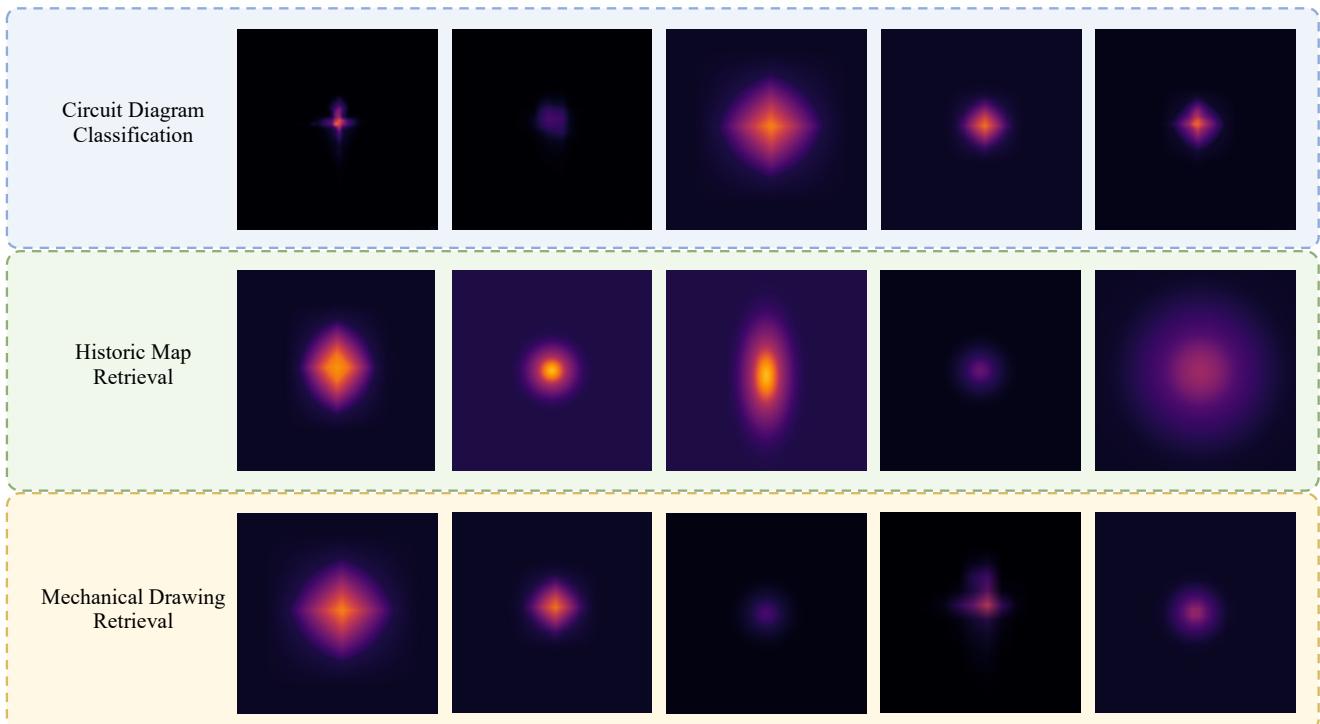


Figure 13. Attention for Specialized Domains. While the attention maps for circuit diagram classification and mechanical drawing retrieval focus more on vertical and horizontal regions, those for historic maps highlight different local regions and tend to have much larger ‘receptive fields’ than the other two tasks.

images for Jersey.

Task III: Mechanical Drawing Retrieval. We collect 565 pairs of mechanical drawings and 3D rendered images from TraceParts [6] and the other 589 pairs come from GrabCAD [4]. These cover various specialized 3D components including brackets, nuts, gears, hinges, and clamps.

E. Visualizations

Attention Maps from Vision Foundation Model on low-resource images. We visualize the attention maps from vision foundation models without our low-resource baselines in Figure 12. This uses the ImageBind model adapted to our low-resource tasks with AdaptFormer. We show the highest activation on each region across all the attention maps of the middle transformer block. We can observe from the figure

that only a few regions are activated for the three images. This means that the vision foundation model fails to understand the interaction between different image regions which is key for these specialized domains. As a result, the model cannot perform well on low-resource tasks. Thus, proper adaptation is needed for vision foundation models.

Attention for Specialized Domains. In Section 3.3, we introduced our attention for specialized domains. Here, we visualize these learned attentions in Figure 13. Each domain tends to have a particular attention pattern with the different attention maps having different sizes of ‘receptive field’ so that various levels of features can be encoded. We observe that the attention for circuit diagram classification and mechanical drawing retrieval focuses more on vertical and horizontal regions. This is because these images have

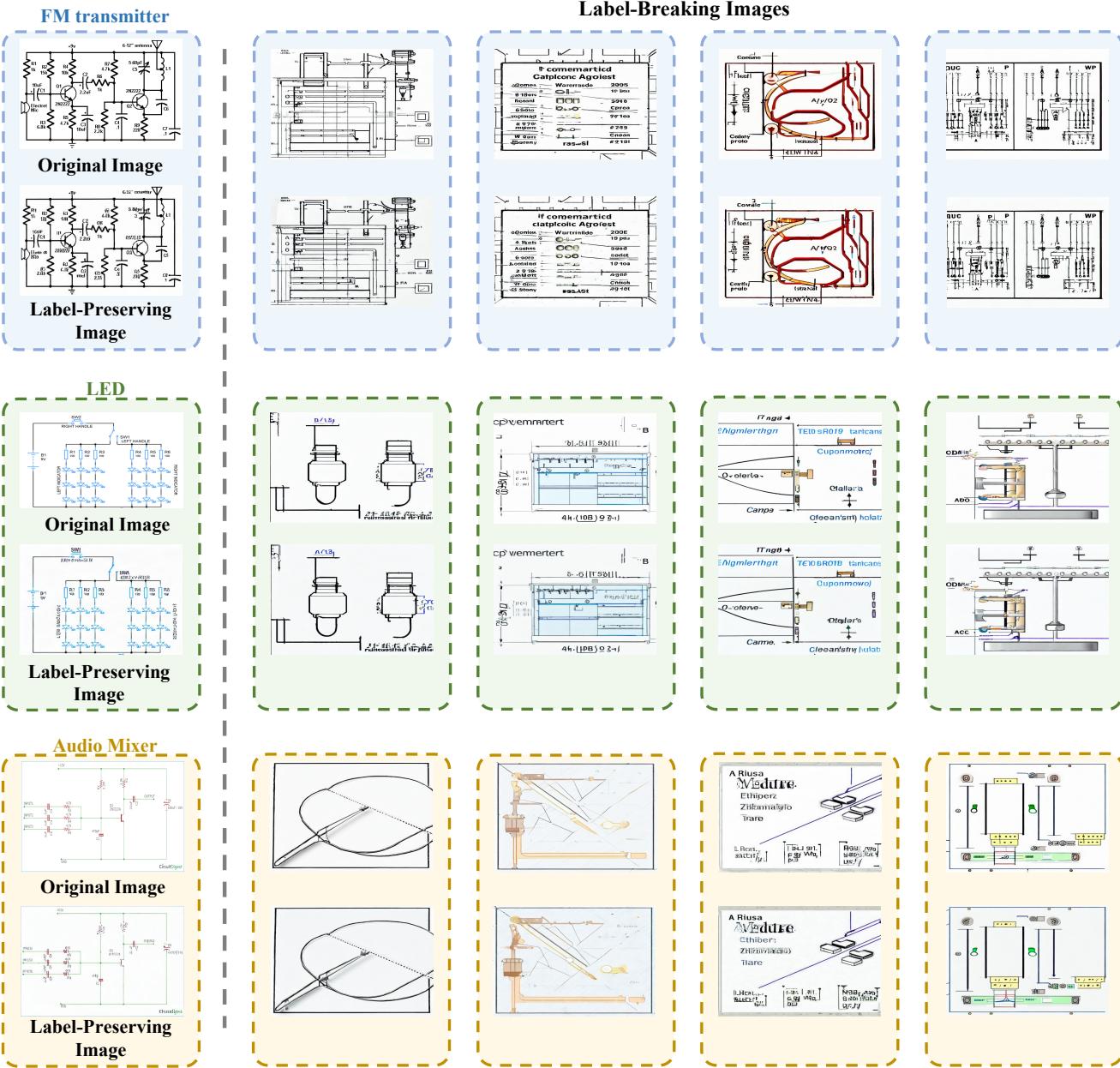


Figure 14. Generated Data of Circuit Diagrams. We show the generated data of three circuit diagram images with different colors. In the left column, we present the original images and their label-preserving augmentations. In the other columns, we show the label-breaking images and the positive pairs for contrastive learning. For the images in each color box, there are minor differences between each other (zoom in to observe the differences), while the label-breaking images are totally different from their original images and have large variations.

a lot of straight lines and right angles, which contain useful information. The attention maps for historic map retrieval highlight different local regions and tend to have much larger ‘receptive fields’ than the other two tasks.

Label-Preserving and Label-Breaking Images. In Section 3.1, we introduce our generated data for data scarcity. Here, we visualize the label-preserving and label-breaking images in Figure 14, Figure 15, Figure 16 and Figure 17. For label-preserving augmentations, there are only minor

differences from their original images (zoom in to observe the differences). Thus, we use the original labels for them in model learning. However, for label-breaking augmentations, they are totally different from the original images. Therefore, we apply a self-supervised contrastive learning objective to learn from such data. We also present the augmentations for label-breaking images, which construct the positive pairs in contrastive learning. Note that each box in the figures denote a positive pair.

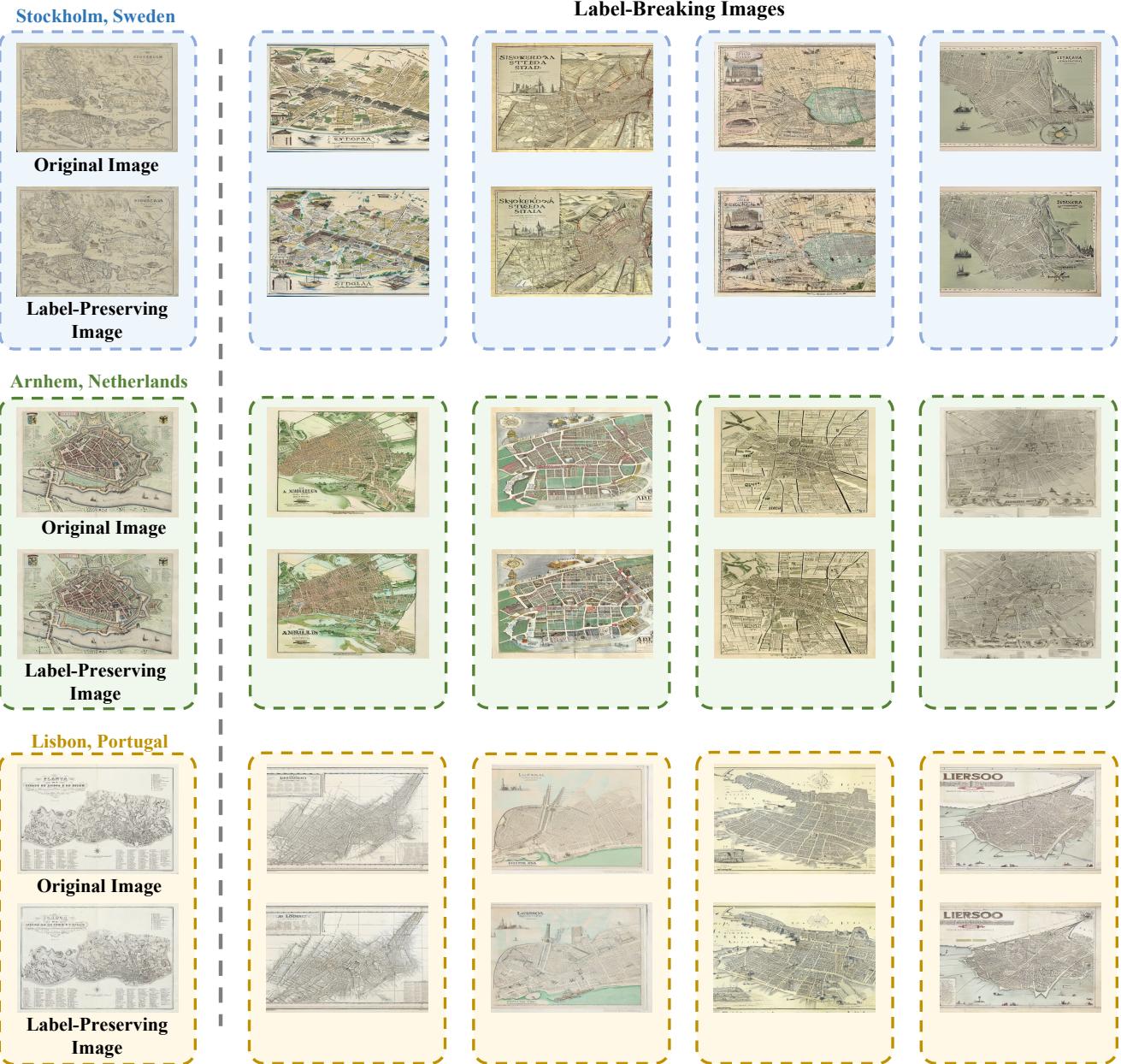


Figure 15. Generated Data of Historic Maps. We show the generated data of three historic maps with different colors. In the left column, we present the original historic map with their label-preserving augmentations. In the other columns, we show the label-breaking images as well as the positive pairs for contrastive learning. For the images in each color box, there are minor differences between each other (zoom in to observe the differences), while the label-breaking images are totally different from their original images and have large variations.



Figure 16. Generated Data of 3D Rendered Images on Mechanical Drawing Retrieval. We show the generated data of three 3D rendered images with different colors. In the left column, we present the original 3D rendered images with their label-preserving augmentations. In the other columns, we show the label-breaking images as well as the positive pairs for contrastive learning. For the images in each color box, there are minor differences between each other (zoom in to observe the differences), while the label-breaking images are totally different from their original images and have large variations.

Label-Breaking Images

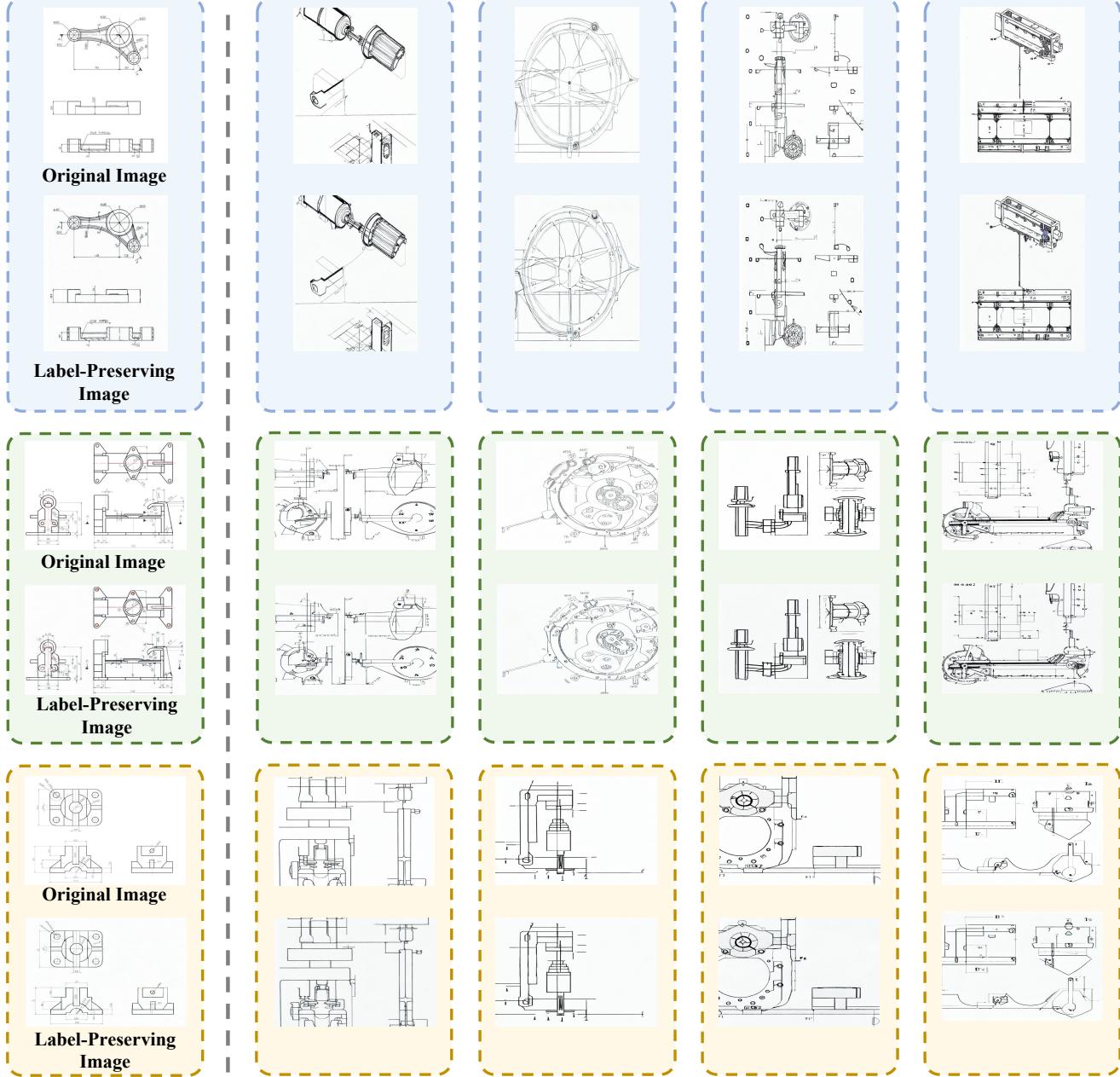


Figure 17. Generated Data of Mechanical Drawings. We show the generated data of three mechanical drawings with different colors. In the left column, we present the original mechanical drawings with their label-preserving augmentations. In the other columns, we show the label-breaking images as well as the positive pairs for contrastive learning. For the images in each color box, there are minor differences between each other (zoom in to observe the differences), while the label-breaking images are totally different from their original images and have large variations.