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 [36]								
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 [16]								
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 [36] and AdaptFormer [16] 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 [82]	20.8	46.0	28.4	55.7	15.2	50.5	72.9	14.9
CutMix [81]	20.0	45.5	28.4	62.6	12.5	54.9	78.1	13.1
Random Erasing [87]	20.8	46.2	23.5	49.6	17.1	54.6	78.0	13.2
Generative Models								
DA-Fusion [65]	19.6	45.1	29.8	61.3	14.7	54.5	77.6	13.5
SyntheticData [31]	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 [14]	16.7	43.2	5.3	12.6	55.1	5.6	17.1	70.2
PLEor [71]	17.1	44.1	4.6	11.8	56.5	5.0	16.9	71.3
PDiscoNet [66]	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				
	Top-1 (%) ↑	Top-5 (%) ↑	R@1 ↑	R@5 ↑	MnR ↓	R@1 ↑	R@5 ↑	MnR ↓	GFLOPs	Params (M)
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 [61]	16.4	43.3	4.2	11.5	52.4	4.4	16.3	69.0	476.2	73.8
CLIP-Adapter [25]	16.3	42.9	3.2	15.9	42.1	7.7	23.2	59.5	224.8	64.2
IA3 [52]	18.2	45.4	29.1	51.3	19.5	52.0	76.7	12.7	224.6	63.6
VPT [41]	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 [36]	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 [16]	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 [14, 66, 71] 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 drawing 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 [16, 25, 36, 41, 52, 61] 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.

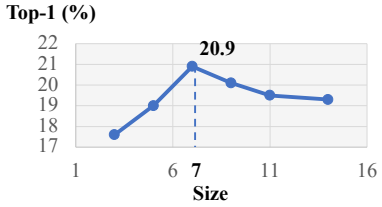

Figure 7. **Effect of Sub-Kernel Size** in tokenization for fine-grained. A medium sub-kernel gives a good trade-off between meaningful regions and fine-grained details.

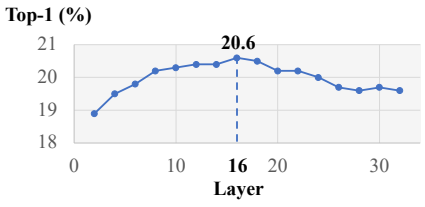

Figure 8. **Position of Attention** for specialized domains. The attention is reasonably robust to the choice of layer, although the middle layers reach a good trade-off between low- and high-level features.

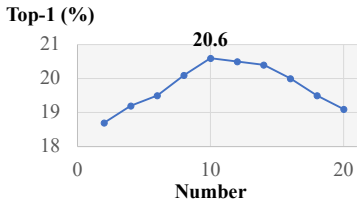

Figure 9. **Number of Attention Maps** for specialized domains. Learning anywhere from 8 to 14 attention maps allows the model to specialize to the domain while avoiding overfitting

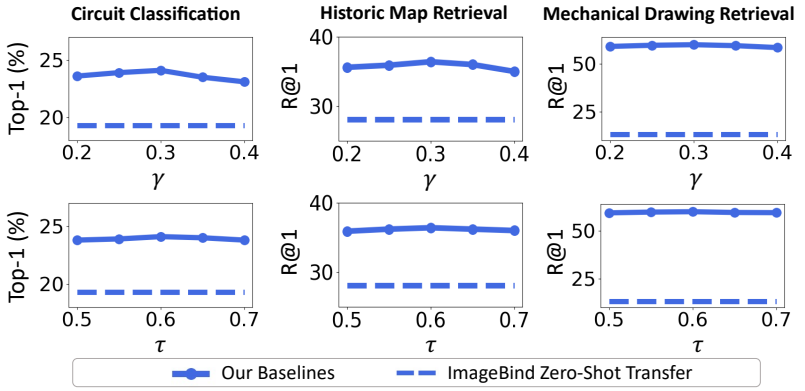

Figure 10. **Diffusion Model Thresholds** have limited influence on baseline I for synthesizing data.

baselines can be used in combination with any transfer learning method we plug them into two such methods: LoRA [36] and AdaptFormer [16]. 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.

Diffusion Model Threshold. Our generated data baseline is insensitive to the selection of γ and τ thresholds and improves over ImageBind zero-shot transfer for all thresholds tested in Figure 10.

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

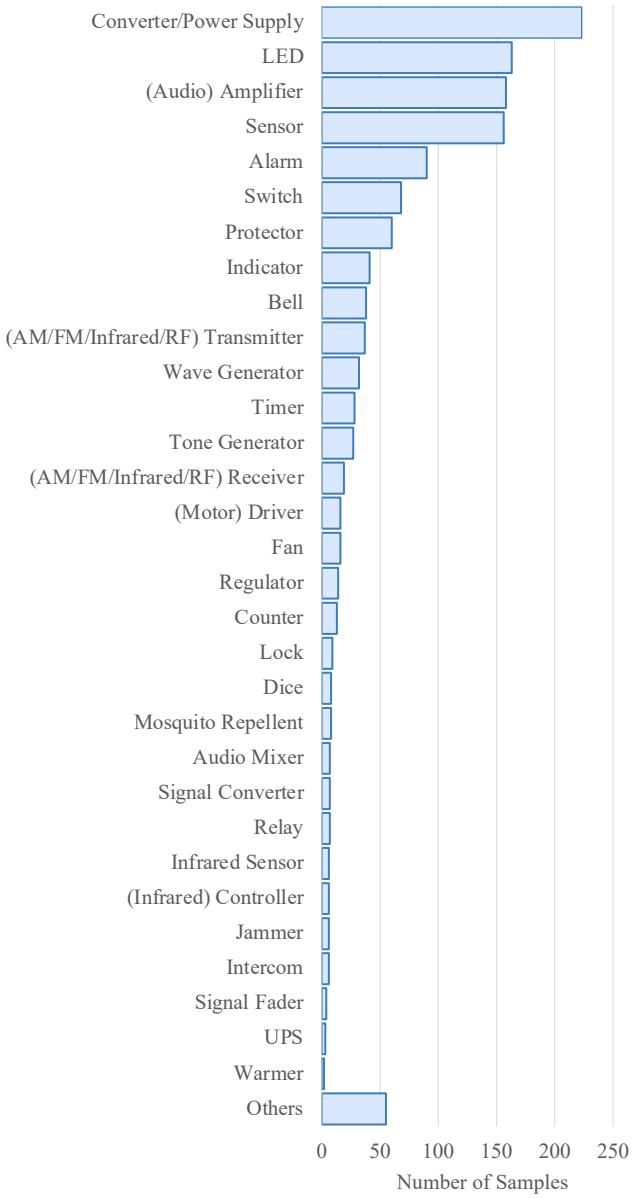

Figure 11. **Class Distribution of Circuit Diagrams**. While power supply, LED, amplifier and sensor have the most samples, it is hard to collect many circuit diagrams for most classes.

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.

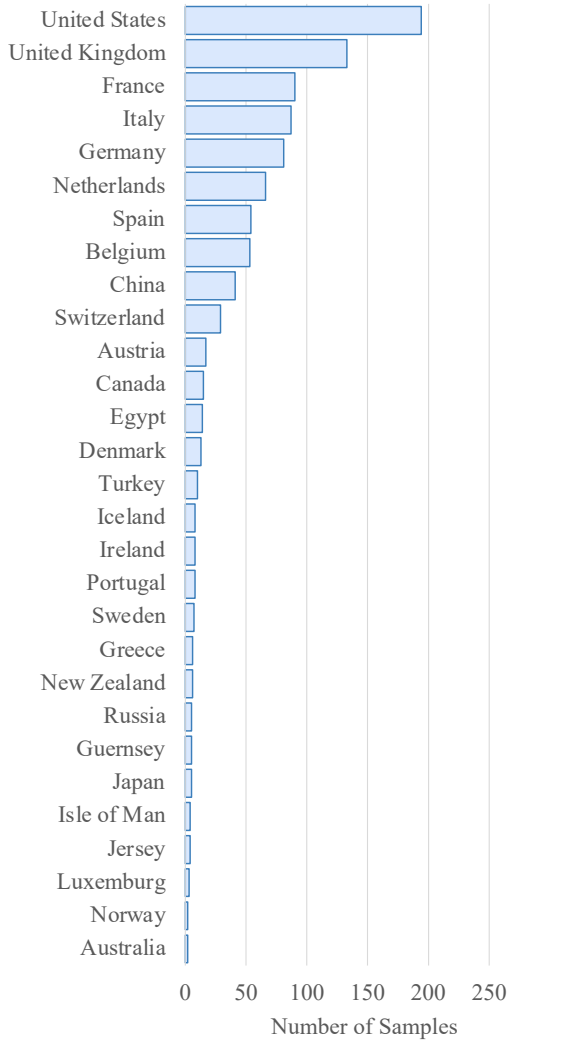

Figure 12. **Country Distribution of Historic Maps**. While we find Europe and United States have the most historic maps online, other regions have much fewer maps available.

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

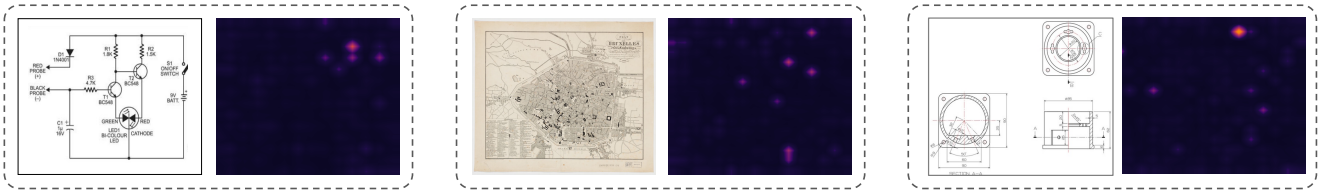

Figure 13. **Attention Maps from Vision Foundation Model**. We show the highest activation on each region across all the attention maps of the middle transformer block. Only a few regions are activated for the three images. This means that the vision foundation model fails to understand the interaction between different image regions. Thus, vision foundation models need proper adaptation for low-resource tasks.

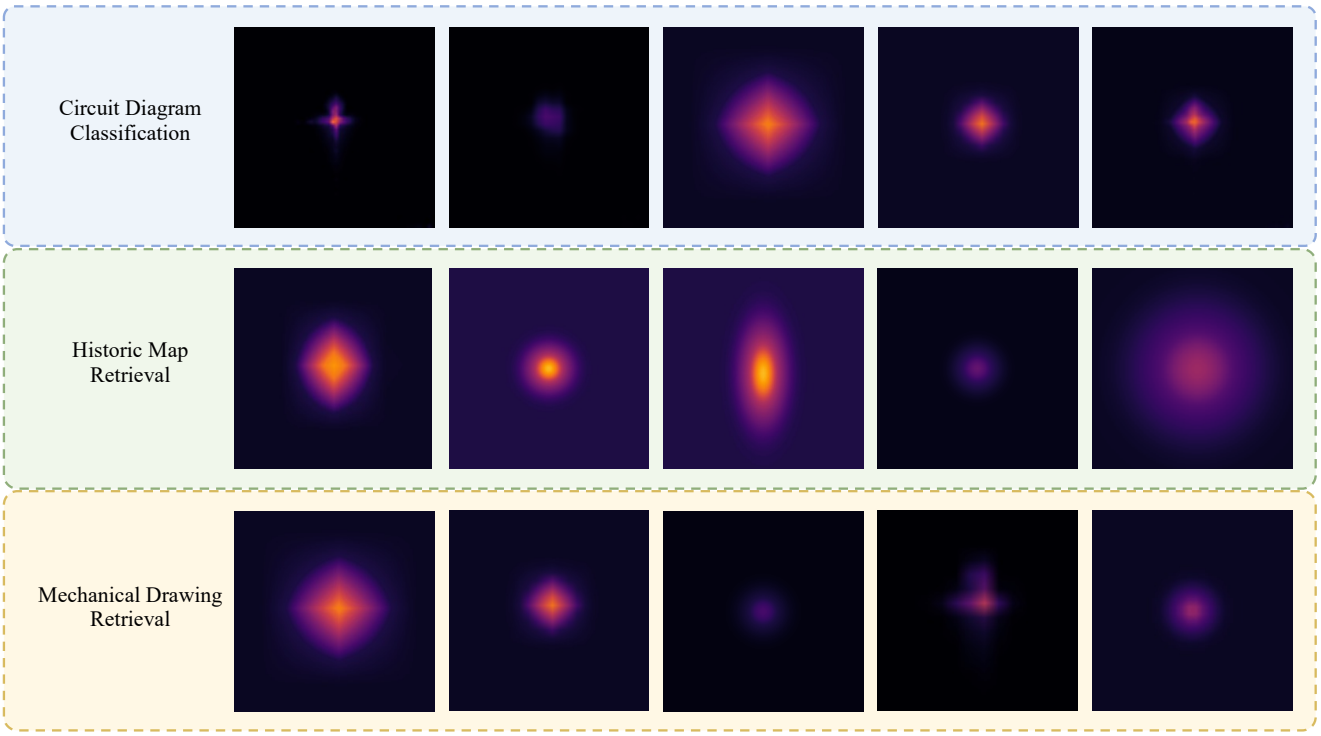

Figure 14. **Attention for Specialized Domains**. While the attention maps for circuit diagram classification and mechanical drawing retrieval focus more on vertical and horizontal regions, those for historic maps highlight different local regions and tend to have much larger 'receptive fields' than the other two tasks.

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. 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 [19], 175 from Gadgetronicx [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 the class distribution in Figure 11. 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 [6] and all the satellite maps are from Google Map [4]. We collect 651 pairs of historic maps and today's satellite images from 29 countries and show the distribution across countries in Figure 12. United States and Europe have the most historic maps online. For

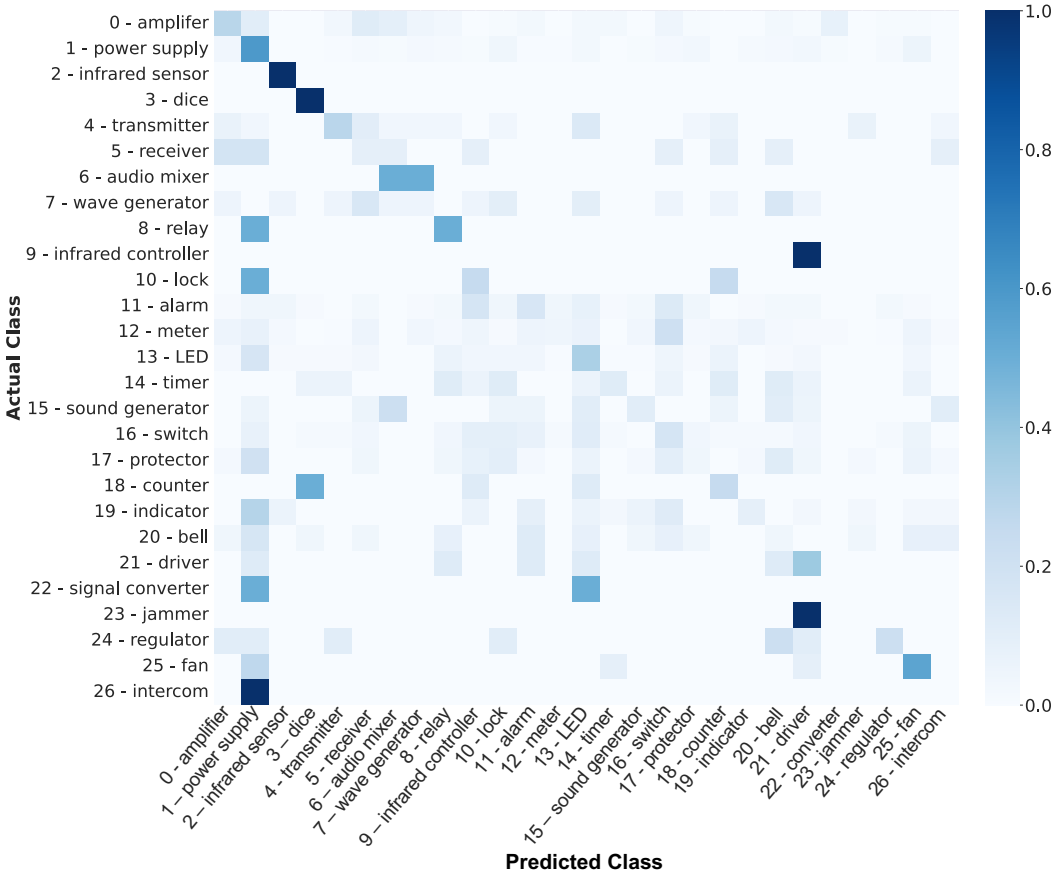

Figure 15. **Confusion Matrix** reveals our baselines achieve stronger performance on classes with prominent patterns.

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 images for Jersey.

Task III: Mechanical Drawing Retrieval. We collect 565 pairs of mechanical drawings and 3D rendered images from TraceParts [7] and the other 589 pairs come from GrabCAD [5]. 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 13. 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 14. 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 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.

Confusion Matrix. We provide a confusion matrix in Figure 15. Indeed, our baselines recognize components with prominent patterns more clearly, achieving stronger performance on *dice*, *infrared sensor*, and *relay*. It confuses classes with shared components, *e.g.*, *audio mixer* and *wave generator*.

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

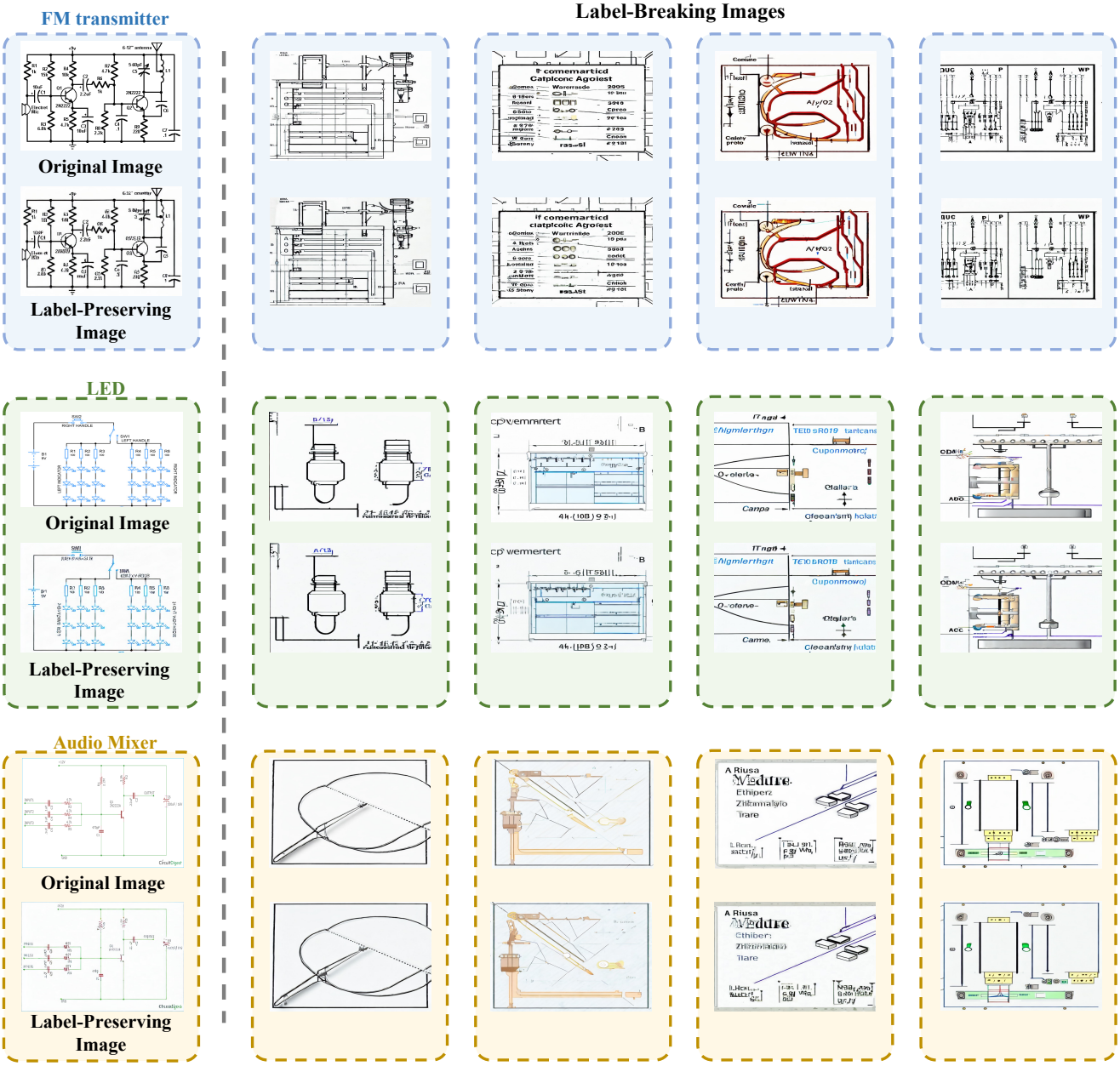

Figure 16. **Generated Data of Circuit Diagrams**. We show the generated data of three circuit diagram images with different colors. In the left column, we present the original images and their label-preserving augmentations. In the other columns, we show the label-breaking images and the positive pairs 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.

images in Figure 16, Figure 17, Figure 18 and Figure 19. 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.

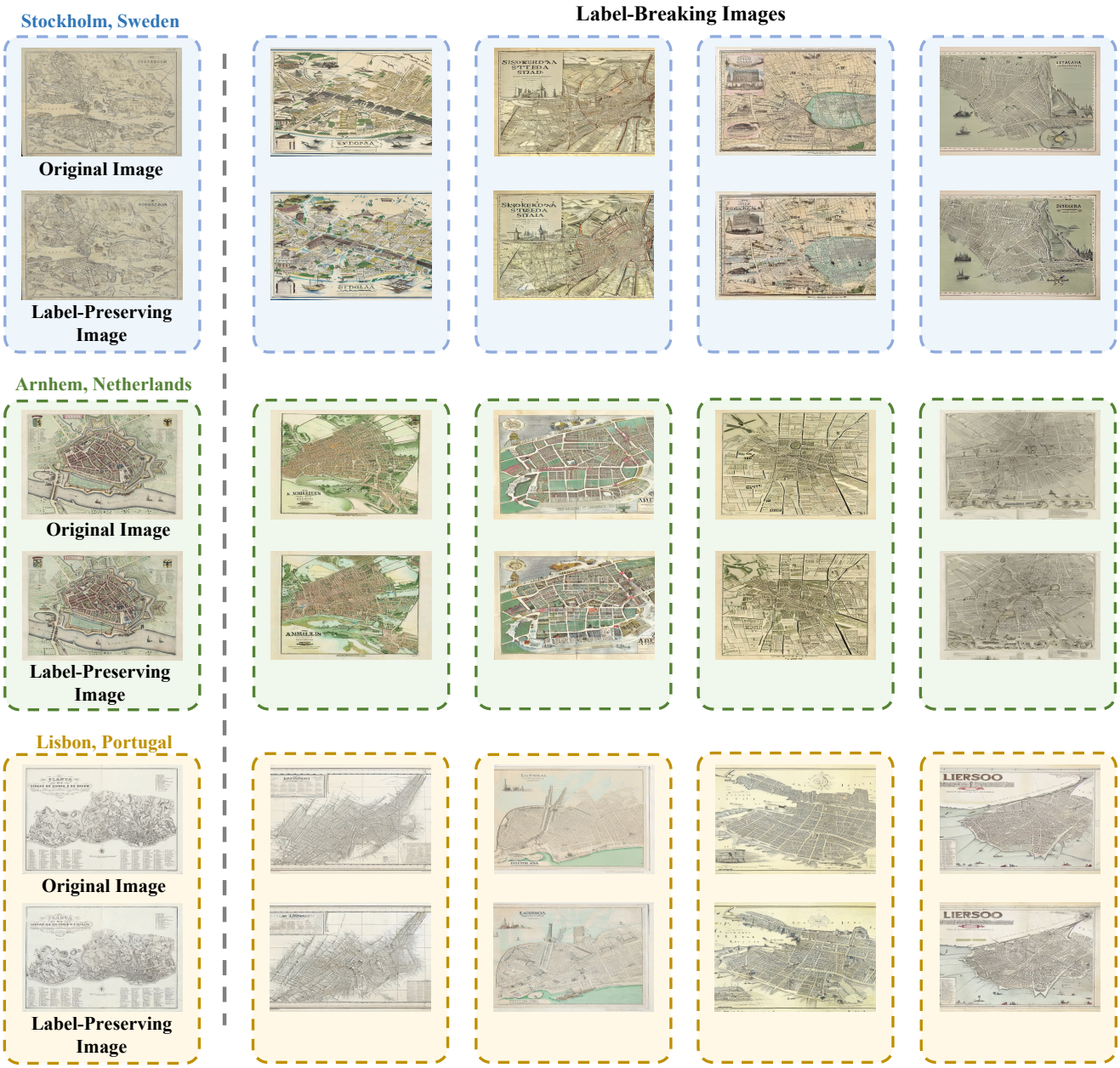

Figure 17. **Generated Data of Historic Maps**. We show the generated data of three historic maps with different colors. In the left column, we present the original historic map with their label-preserving augmentations. In the other columns, we show the label-breaking images as well as the positive 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 18. **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.

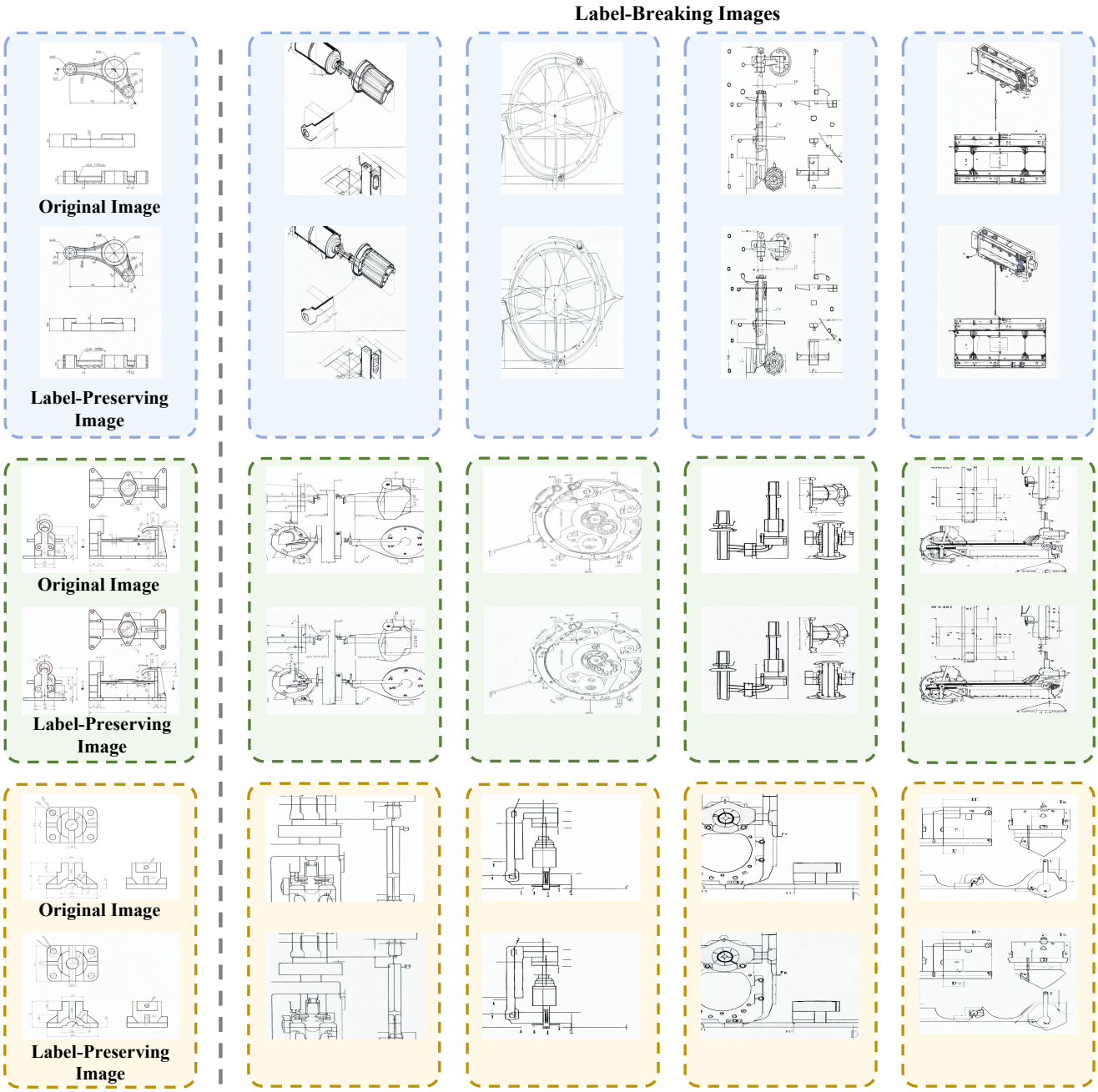

Figure 19. **Generated Data of Mechanical Drawings**. We show the generated data of three mechanical drawings with different colors. In the left column, we present the original mechanical drawings with their label-preserving augmentations. In the other columns, we show the label-breaking images as well 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.