TouhouIC: Accurate Image Classifier at Minimal Cost with Transfer Learning and Data Augmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

This project explores cost-efficient strategies for building accurate image classifiers, with an focus on noisy, domain-specific data. We construct a Touhou Project character dataset comprising 120 categories, each with approximately 1,100 images, of which only one-tenth are manually annotated as test set. Various models and techniques are evaluated to balance accuracy and resource efficiency. A fine-tuned ViT-large model achieves the best baseline accuracy of 92% within 10 training epochs, demonstrating the effectiveness of transfer learning. With data augmentation, we further improves performance to 96%, also using ViT-large, without requiring additional labeled data or significantly increased computational cost. Our findings highlight the benefits of transfer learning and data augmentation for image classification at minimal cost.

1 Introduction

Image classification (IC) has always been a foundational task in computer vision, powering various applications. Over the past decade, deep learning has significantly advanced the state of the art in IC, with convolutional neural networks (CNNs) leading the way for years(11; 10; 19; 5). More recently, Vision Transformers (ViTs) have emerged as a powerful alternative, leveraging self-attention mechanisms and pretraining on large-scale image datasets to exceed CNN performance(3; 4).

Despite these advancements, real-world application of modern IC techniques to specific domains often faces two major obstacles: high computational requirements and the prohibitive cost of manual data annotation. While large organizations can afford multi-GPU clusters and massive labeled datasets, many users—especially in niche or enthusiast domains—do not have access to these resources. For example, someone interested in classifying characters from the Touhou Project(23), a long-running Japanese multimedia franchise, may struggle to collect clean labeled data or train large models due to limited hardware and budget.

To address these challenges, transfer learning has become a key strategy (25; 20). Large models such as ViTs, pretrained on large-scale datasets like ImageNet, acquire the ability to capture generic, low-level visual features that are broadly applicable in various domains. This enables them to be adapted to different downstream tasks with substantially reduced data and computation requirements, manifesting remarkable generalization ability.

Complementing this, data augmentation enhances training by generating synthetic variations of existing images through transformations like flipping, cropping, and color jittering (18; 26). This reduces overfitting and improves performance, particularly when annotated data is scarce.

This project conducts an empirical study combining transfer learning and data augmentation to classify Touhou Project characters. We explore a range of model-strategy combinations and implement the complete workflow—from dataset collection to model selection, training, and evaluation. Our results

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

reaffirm the effectiveness of transfer learning and data augmentation in improving accuracy under limited data and resource constraints.

2 Background and Related Work

2.1 Image Classification

Image classification is a core task in computer vision that categorizes images into predefined groups. It serves as the foundation for numerous real-world applications, throughout medical diagnostics, content filtering, autonomous vehicles, facial recognition systems to agriculture production monitoring (12; 21; 16).

Early approaches to image classification relied on manual feature extraction techniques, such as Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG), combined with traditional classifiers like Support Vector Machines (SVMs) (14; 1). These methods required significant domain expertise and often struggled with complex visual patterns.

The introduction of Convolutional Neural Networks (CNNs) marked a significant advancement in image classification. LeNet-5, demonstrated the potential of CNNs in recognizing handwritten digits (11). Subsequent architectures, such as AlexNet (10), VGGNet (19), and ResNet (5), further improved performance by increasing network depth and introducing novel architectural components.

In parallel with architectural advancements, the increasing availability of large-scale datasets such as ImageNet (2) significantly accelerated progress in image classification. With over a million labeled images spanning thousands of categories, ImageNet enabled the effective training of larger deep neural networks, yielding substantial performance improvements. These developments cemented CNNs as the dominant approach throughout the 2010s and paved the way for further advances.

2.2 Transfer Learning

Transfer learning has emerged as a practical and effective approach to address challenges in training deep neural networks with limited data and computational resources(20). This technique involves leveraging knowledge from a model pre-trained on a large source dataset and adapting it to a target task with smaller-scale data, demonstrating promising performance across various machine learning domains including natural language processing, text-to-speech synthesis, and image generation

As a key implementation strategy, fine-tuning adjusts parameters of pre-trained models on new datasets through either full-parameter adaptation or partial-parameter optimization such as LoRA(6). Both approaches significantly enhance model performance while requiring minimal computational resources, making them particularly suitable for large-scale architectures.

In computer vision applications, transfer learning typically involves fine-tuning convolutional neural networks (CNNs) on specialized datasets. This process enables retention of generalized feature extractors while adapting classification heads to novel categories. The methodology has achieved remarkable success in medical imaging analysis and domain-specific object recognition(8).

2.3 Data Augmentation

Data augmentation is a strategy employed to artificially expand the size and diversity of training datasets, thereby enhancing model generalization and mitigating overfitting. Common augmentation techniques include geometric transformations (rotating, scaling, and flipping images), photometric adjustments, and more advanced methods. These techniques introduce variability in the training data, enabling models to become more robust to variations in input images. Data augmentation is particularly beneficial in scenarios with limited labeled data, as it effectively increases the dataset size without additional manual annotation.

Photometric Adjustments involve modifications to the color properties of images. Techniques such as adjusting brightness and contrast can help the model become more robust to changes in lighting conditions. Color jittering (26), which includes random adjustments to hue, saturation, and intensity, can simulate different environmental lighting conditions and camera sensitivities, thus preparing the model for a variety of real-world scenarios.

Advanced augmentation methods like CutMix and MixUp (24) introduce more complexity by blending parts of two different images or their labels to create new training samples. CutMix involves cutting a rectangular region from one image and pasting it onto another, while MixUp interpolates between pairs of images and their labels. These methods not only increase the diversity of the training data but also help in regularizing the model by preventing it from becoming overly reliant on any single feature.

3 Method

In this section, we present the overall workflow and key components of our approach (see Figure 1).

- We begin by detailing how the training and test datasets were constructed.
- Next, we discuss our model selection process and justify our choice of ViT-large for fine-tuning.
- We then describe the training techniques employed to maximize efficiency under limited compute resources, including mixed-precision, optimizer schedules, and early stopping.
- Following this, we introduce our self-inspection data filtering strategy to mitigate label noise.
- Finally, we outline our multi-stage data augmentation pipeline, comprising diversity- and generalization-oriented transforms as well as MixUp/CutMix, to further enhance model robustness.

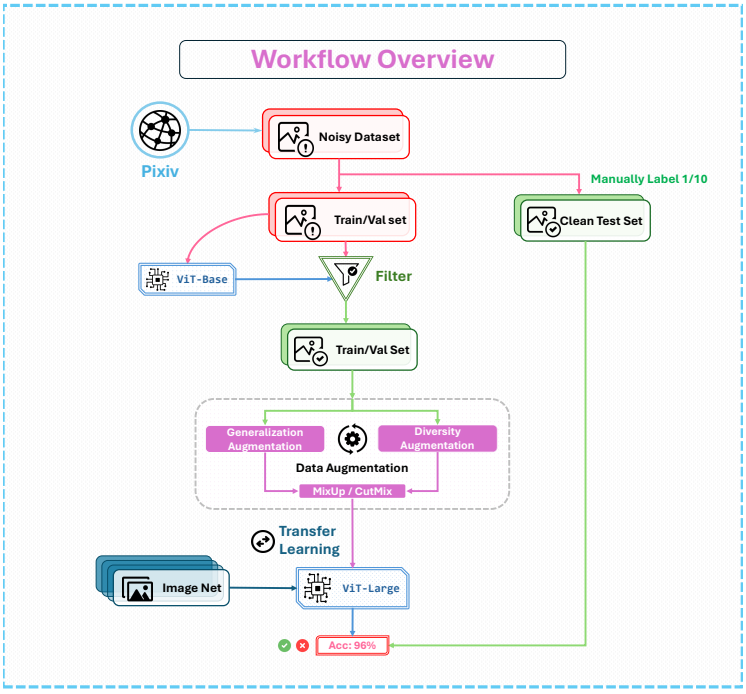

Figure 1: Workflow Overview

3.1 Dataset Construction

Niche-specific image classification tasks often lack readily available datasets. For this study, we sourced our dataset from Pixiv (17), a popular online art community. Pixiv conveniently provides a 256x256 thumbnail for each illustration at search page. We collected the list of Touhou characters from THBWiki (22) as categories. To gather relevant images, we searched Pixiv using the Japanese names corresponding to our desired labels and collected the top 1100 most popular thumbnails for

each. To ensure data quality and sufficient representation, we filtered out categories with fewer than 2000 search results due to lack of high-quality images.

For the test dataset, we randomly sampled 100 images for each label from the initial collection of 1100 images per category. To ensure the quality and consistency, each selected image in test set was manually labeled and verified.

3.2 Model Selection

During the preliminary experiments, we evaluated a variety of image classification architectures, including ResNet, ViT-Base, ViT-Large, and a Mixture-of-Experts (MoE) (7) variant based on ViT-Base. Due to limited GPU memory, our choices were constrained—larger models could not be accommodated, which is a common challenge for niche or resource-constrained projects.

We explored both training from scratch and fine-tuning from pretrained weights. When trained from scratch, only the ResNet-based models showed stable convergence, with performance improving as model capacity increased. In contrast, Vision Transformer variants—including ViT-Base and ViT-Large—suffered from severe overfitting and achieved only around 30% accuracy on the validation set, indicating poor generalization ability in this low-data regime.

However, when fine-tuned from ImageNet-pretrained weights, the performance of ViT-based models improved dramatically. Both ViT-Base, ViT-Large, and the MoE variant significantly outperformed ResNet, demonstrating the effectiveness of transfer learning in overcoming data limitations. Among them, ViT-Large achieved the highest accuracy, making it the most suitable choice for our downstream task.

Therefore, all subsequent experiments are based on fine-tuning the pretrained ViT-Large model.

3.3 Training Techniques

To achieve both time and cost efficient training on limited computation resources, training techniques must be meticulously considered.

First, we adopt mixed precision training to leverage the FP16 capabilities of our NVIDIA Ampere or Ada accelerators. This approach significantly enables larger batch sizes and higher throughput, with only a minimal trade-off in convergence speed. (15)

Specifically, we perform forward and backward computations in FP16, while model parameters are maintained and updated in FP32. Loss scaling is consistently applied to prevent gradient underflow with FP16.

Second, we choose SGD combined with learning rate scheduler for pre-training; AdamW (13; 9) and full model fine-tuning for tuning. To maximize computational throughput, the batch size is set to the largest value the GPU memory can accommodate (e.g., 30 on a 12GB GPU). As the batch size in our case is not large, statistical efficiency issue is negligible. The dataset is partitioned with a $9:1$ ratio for training and validation. Other detailed hyperparameters are presented in Table 1.

Third, regarding early stopping, we halt training if the validation loss does not show improvement for N epochs. Empirically, we set N to 3. Besides, checkpoint is saved every epoch to find desirable data filtering judge with reasonable accuracy and minimal overfitting, as described in the next subsection.

Type	Hyperparameters
Pre-training	Optimizer: SGD Learning Rate: 5×10^{-3} LR Scheduler: StepLR (step_size=5, gamma=0.25)
Fine-tuning	Method: Full Model Fine-tuning Optimizer: AdamW Learning Rate: 1×10^{-5} Weight Decay: 1×10^{-2}

Table 1: Training Hyperparameters for Pre-training and Fine-tuning

3.4 Data Filtering

To address the label noise, we adopt a self-inspection method to filter out the mislabeled data points. Denoting a filter judge:$M : X \to Y$, we construct the filtered dataset D' as follows:

$$D' = \{(x, y) \in D | M(x) = y\} \tag{1}$$

which retain only the entries whose labels are consistent with the judge prediction.

We choose the ViT-base model - finetuned on the original training dataset and early stopped after very few epochs - as the filter judge. The judge model achieves $\sim 88\%$ accuracy on the test set, and has comparable training loss to validation loss. Its high accuracy and low overfitting feature ensure the effectiveness of our filtering method.

Figure 2: Data Filtering Example for Label Cirno

3.5 Data Augmentation

The data augmentation strategy is composed of three components: *Diversity Augmentation* , *Generalization Augmentation* , *MixUp/CutMix*.

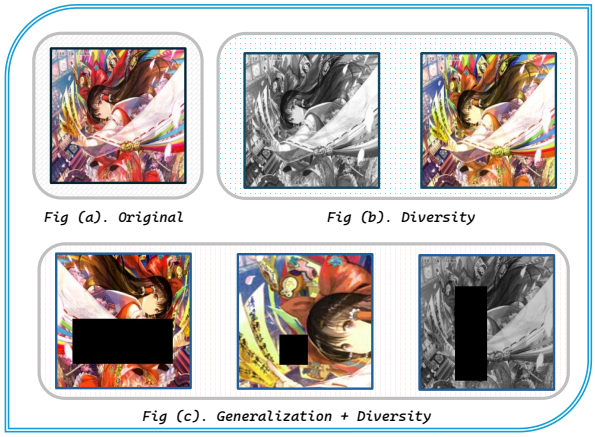

Figure 3: Data Augmentation Example

Diversity Augmentation: The Diversity Augmentation strategy is specifically designed to augment the visual diversity of the training dataset by simulating a variety of artistic styles and illustrators. This is accomplished through two primary techniques: *random grayscale conversion* and *color jittering*. Random grayscale conversion involves the stochastic transformation of RGB images into their grayscale equivalents, thereby increasing the proportion of black-and-white, comic-like images within the training data. These monochromatic images are prevalent in our dataset but often pose greater challenges for model learning. Color jittering enables the random perturbation of image

brightness, contrast, saturation, and hue, thereby capturing the variability in color preferences across different illustrators and simulating a broader spectrum of artistic styles.

Generalization Augmentation: The Generalization Augmentation strategy aims to enhance the model's ability to generalize to unseen data by incorporating a suite of techniques that introduce variability and invariance into the training process. Specifically, it integrates methods such as *random resizing and cropping, random horizontal flipping, and random erasing*. These techniques facilitate the model's learning of more robust and invariant features, thereby reducing its reliance on any single, unique feature for target class identification. Consequently, the model becomes more adept at recognizing target classes under diverse conditions, even in the absence of certain distinctive features, thus improving its overall accuracy and generalization capability.

MixUp/CutMix: Following the application of the aforementioned augmentations, the training process is further refined through the stochastic application of either MixUp or CutMix to each training batch. Specifically, MixUp synthesizes novel samples by constructing convex combinations of paired images and their associated labels. In contrast, CutMix involves substituting a rectangular region of one image with a patch derived from another image, with corresponding adjustments made to the labels. Both techniques markedly enhance the diversity of the training samples and effectively mitigate the influence of intrinsic label noise within the dataset.

It is worth noting that our data augmentation strategy involves randomly selecting a set of augmentations for each instance at each epoch, rather than explicitly expanding the dataset by applying all augmentations beforehand. Therefore, for each epoch, number of samples is the same as the case without data augmentation.

4 Evaluation

We evaluate our method on the Touhou character test set described in subsection 3.1, with 120 categories and 1000 training/validation and 100 test samples per category.

4.1 Accuracy Results

We test six settings listed in Table 2:

Model	Dataset	Data Augmentation	Training Method
ViT-Large	Filtered	Yes	Finetuned
ViT-Large	Raw	Yes	Finetuned
ViT-Large	Filtered	No	Finetuned
ViT-Large	Raw	No	Finetuned
ViT-Base	Raw	No	Finetuned
ResNet-152	Raw	No	Pretrained

Table 2: Evaluation Settings

Note that both ViT-Base or ViT-large tends to overfit rapidly at pretraining, leading to a poor accuracy around 40%, so we instead choose ResNet-152 that has the best performance with pretraining as our baseline.

The results is shown in Fig 4.

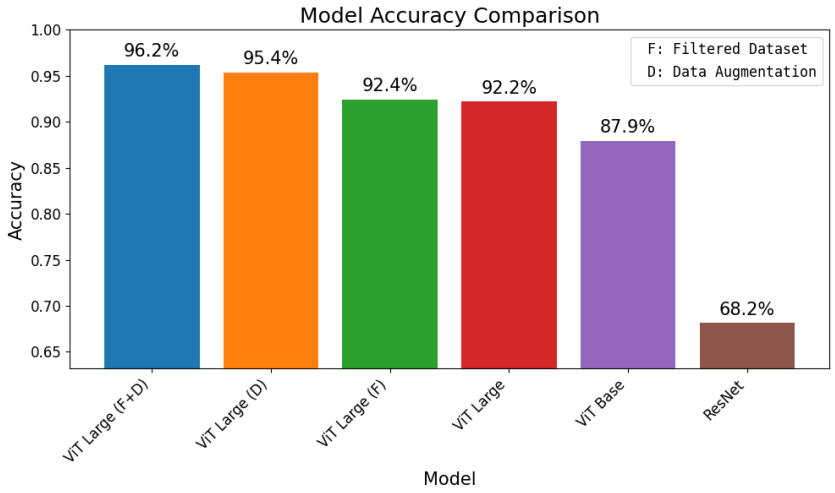

Figure 4: Accuracy Comparison

Our analysis reveals a significant performance uplift for the ViT-Large model when both data filtering (F) and data augmentation (D) are applied. Specifically, ViT-Large(F+D) achieves the highest observed accuracy of 96.2%. This represents a substantial improvement of 28.0% over the baseline ResNet-152 model that utilize none of F, D or transfer learning.

Validation of transfer learning:

Notably, the ViT-Large model, even without any data filtering or augmentation, already achieves a significant 19.7% accuracy improvement over the best pre-trained ResNet model. This substantial gain highlights the intrinsic power and superior representational capabilities of the Vision Transformer architecture, demonstrating its strong foundational performance even prior to dataset-specific enhancements.

Since ViT-Large is observed to fail in generalizing when pre-trained solely on our relatively small datasets, transfer learning proves essential for enabling more powerful large model on such a specific, data-scarce task.

Validation of data augmentation:

ViT-Large (F+D) gains an improvement of 4.0% compared to that applies neither filtering nor data augmentation. This marks combined effectiveness and synergistic power of employing both data filtering and data augmentation strategies.

Delving deeper into the individual contributions, we find that data augmentation alone, even without any filtering, significantly boosts accuracy by 3.2%. This highlights the robust capability of data augmentation to enhance model generalization by expanding the diversity of the training data.

Interestingly, while data filtering on its own yields only a marginal 0.2% improvement in accuracy, its true value becomes apparent when integrated with data augmentation. In this combined scenario, filtering contributes an additional 0.8% improvement on top of what data augmentation provides alone, leading to the overall 4.0% gain from the baseline. This suggests that while filtering might not directly add substantial information, it likely plays a crucial role in preparing or refining the dataset such that the benefits of data augmentation are amplified, perhaps by removing noisy samples that would otherwise dilute the impact of augmented data.

4.2 Ablation Study of Data Augmentation and Filtering

To evaluate the contribution of each data augmentation/filtering component, we conduct a systematic ablation study using the ViT-Large model finetuned on our filtered dataset as the default setting. We selectively disable individual components from the full pipeline and compare the resulting performance on the test set. This allows us to quantify the impact of each module – *Filtering*,

Generalization Augmentation, *Diversity Augmentation*, and *MixUp/CutMix* – in addition to disabling all augmentations or (filtering + all augmentations).

Setting	Accuracy (%)	Difference (%)
Filtering + All Augmentations	**96.2**	0
- w/o Filtering + All Augmentations	92.2	-4.0
- w/o All Augmentations	92.4	-3.8
- w/o Filtering	95.4	-0.8
- w/o Generalization Augmentation	95.9	-0.3
- w/o Diversity Augmentation	96.1	-0.1
- w/o MixUp/CutMix	95.4	-0.8

Table 3: Ablation Study Results: Removing individual components from the training pipeline.

From Table 3, we observe that removing filtering or MixUp/CutMix alone results in notable performance degradation of 0.8%, suggesting orthogonality between these two methods. This contradicts the expectation that MixUp/CutMix would smooth label noisy to achieve similar effect as filtering.

Removing generalization or diversity augmentation alone causes only marginal accuracy loss ($0.1\% \sim 0.3\%$).

In general, disabling individual augmentation component ($0.1\% \sim 0.8\%$) does not result in comparable accuracy drop regarding w/o All Augmentations (3.8%), demonstrating complementary feature of each augmentation methods.

Regardless, each component does increase some accuracy. Given that adding each component does not incur additional computation cost, combining all available data augmentation methods and filtering is beneficial.

4.3 Training Cost

We utilize three commodity GPUs that are commonly available for personal desktops or laptops for experiment:

- NVIDIA RTX 4070 Ti Super (285W, 16GB)
- NVIDIA RTX 4070 Super (220W, 12GB)
- NVIDIA RTX 3080 Ultra (350W, 12GB)

These GPUs have similar computation performance, so we only report statistics for NVIDIA RTX3080 Ultra.

In our training setup described in subsection 3.3, model converges within 20 epochs under all finetuning setting. Each epoch takes around 30 minutes to iterate through 120k samples. The total training cost for one setting is bounded by 10 hours and 2.4 million samples.

Transfer Learning:

The ViT-Large/32 pretraining processes roughly 2.9 billion samples—comprising 300 epochs on ImageNet-1K, 30 epochs on ImageNet-21K, and 7 epochs on JFT-300M (3)—which is three orders of magnitude greater than our fine-tuning setup (2.4 million samples).

This stark contrast highlights the **training efficiency of transfer learning**. If we were to train a comparably performant model from scratch, it would require not only massive computational resources but also access to large-scale, high-quality datasets. In contrast, leveraging pretrained models significantly reduces the training cost while still achieving strong performance, making it a practical solution for limited-resource scenarios.

Data Augmentation:

Since our data augmentation persists the number of samples processed per epoch, and convergence speed is similar, applying data augmentation does not incur additional training cost.

Mixed Precision Training:

Mixed precision training profoundly speeds up the training by 2x. It reduces per epoch time from over 60 minutes to 30 minutes by enabling larger batch size and better GPU utilization.

Moreover, it is also possible to employ our training method on lower-end GPUs, making it more accessible to a wider range of hardware configurations.

4.4 Dataset and Labeling Cost

The construction of our dataset is designed to minimize manual labeling effort:

- **Data fetching:** Crawling hundreds of thousands of the 256×256 thumbnails from Pixiv requires several hours of automated scraping, but incurred no manual labor cost.
- **Labeling effort:** We manually label only the test set (100 images per category) to verify model performance. This verification step is optional and not required for model training. One may rely solely on noisy dataset and subsequently filtered dataset provided that noise is not too great.
- **Data augmentation:** On-the-fly augmentation leverages existing images without human intervention.

In summary, our approach can achieve competitive finetuning results with virtually zero labeling cost beyond optional test-set verification.

5 Discussion

Our method has demonstrated a high level of accuracy on the niche dataset, highlighting the effectiveness of combining transfer learning, data filtering, and data augmentation in specialized image classification tasks.

With 307 million parameters, ViT-Large exhibits strong transfer learning capabilities even when applied to relatively small-scale datasets. This result suggests that large vision models pretrained on general-purpose datasets such as ImageNet retain sufficient feature abstraction capacity to adapt effectively to niche domains like the Touhou Project, where labeled data is limited and highly variable in artistic style.

The data filtering mechanism plays a crucial role in improving dataset quality by removing mislabeled or noisy samples while preserving correctly labeled and semantically similar images. We use ViT-Base—fine-tuned for only a few epochs—as a lightweight filter judge to identify inconsistent labels. As shown in Figure 2, many of the filtered-out positive samples exhibit high similarity with their correct counterparts, indicating that they may represent edge cases rather than clearly incorrect labels. These ambiguous samples often contain shared visual features that are still informative for training. On the other hand, negative samples removed during filtering tend to be particularly challenging instances, such as black-and-white comic-like images or those with incomplete character depictions. While these samples are excluded from the clean dataset, their absence is effectively mitigated by the subsequent application of data augmentation techniques.

Data augmentation contributes significantly to model performance improvement, serving three key purposes: (1) enriching the dataset with difficult or underrepresented samples, (2) enhancing the model's generalization ability, and (3) smoothing label noise. These correspond respectively to the Diversity Augmentation, Generalization Augmentation, and MixUp/CutMix components. Specifically, Diversity Augmentation introduces stylistic variation through color jittering and random grayscale conversion, mimicking different illustrators' styles and thereby increasing the diversity of hard-to-classify samples. This component must be carefully designed for niche domains, where artistic variability is high and standard augmentation strategies may not fully capture the target distribution.

In summary, our empirical analysis highlights the importance of combining transfer learning with effective data curation and augmentation strategies when working with limited and noisy data in niche domains. The results suggest that such an approach can achieve high classification accuracy without requiring extensive labeled datasets or computational resources.

Limitations

Despite the strong performance of our approach, there are several limitations that should be acknowledged.

First, our data filtering mechanism relies on weakly labeled data for training the filter judge model, which inherently assumes that the initial dataset is mostly correct. This limits the applicability of our method in scenarios where label noise is pervasive or where high-quality seed labels are unavailable.

Second, while the filtering process effectively removes mislabeled or noisy samples, it also discards potentially useful images that could contribute to training. This leads to a reduction in dataset size and may inadvertently exclude rare or challenging but valid examples, especially in domains with already limited data diversity.

Lastly, the current diversity augmentation strategy is manually designed and tailored specifically for the Touhou Project dataset. As a non-parametric method, it lacks adaptability across different artistic styles or domains, limiting its generalization potential. Future work could explore more flexible, parametric augmentation techniques—such as style-aware generative models or automated augmentation policies—to improve cross-domain transferability.

6 Conclusion

In this work, we presented an empirical study on image classification within a niche domain—the Touhou Project—demonstrating that high-performance results can be achieved with minimal computational resources and limited weak labeled data. By combining transfer learning with effective data filtering and carefully designed data augmentation strategies, we significantly improved classification accuracy over baseline models, achieving up to 96.20% test accuracy using a fine-tuned ViT-Large model.

Our approach reduces the reliance on large-scale annotated datasets and high-end computing hardware, making it particularly suitable for small-scale or enthusiast-driven domains where such resources are often unavailable. The proposed data filtering mechanism effectively mitigates label noise, while the two-stage augmentation strategy—balancing diversity and generalization—enhances model robustness and adaptability.

For future work, we plan to extend this framework to other niche domains with similar constraints, such as character classification in less well-known anime or fan-made art styles. Additionally, exploring more sophisticated augmentation techniques—both parametric (e.g., GAN-based style transfer) and non-parametric (e.g., adaptive MixUp/CutMix sampling)—could further improve performance. Investigating automatic augmentation policy search (e.g., AutoAugment-style methods) tailored to artistic styles is also a promising direction.

Overall, our findings reaffirm the power of transfer learning and data-efficient training strategies in enabling high-quality image classification under resource-constrained settings.

References

[1] DALAL, N., AND TRIGGS, B. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (2005), vol. 1, IEEE, pp. 886–893.

[2] DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., AND FEI-FEI, L. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition* (2009), 248–255.

[3] DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., HOULSBY, N., KOWALSKI, M., SCHMID, G., GELLY, S., ET AL. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021).

[4] HAN, K., XIAO, A., DING, E., XU, G.-J., LI, C., LI, S., AND DING, C. A survey of vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 10 (2022), 7012–7027.

[5] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778.

[6] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021.

[7] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural computation 3*, 1 (1991), 79–87.

[8] Kandel, I., and Castelli, M. How deeply to fine-tune a convolutional neural network: A case study using a histopathology dataset. *Applied Sciences 10*, 10 (2020).

[9] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization, 2017.

[10] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), vol. 25, pp. 1097–1105.

[11] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[12] Li, M., Jiang, Y., Zhang, Y., and Zhu, H. Medical image analysis using deep learning algorithms. *Frontiers in Public Health Volume 11 - 2023* (2023).

[13] Loshchilov, I., and Hutter, F. Decoupled weight decay regularization, 2019.

[14] Lowe, D. G. Distinctive image features from scale-invariant keypoints. In *International journal of computer vision* (2004), vol. 60, Springer, pp. 91–110.

[15] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017).

[16] OpenCV. Image classification in 2025: Insights and applications. `https://opencv.org/blog/image-classification/`, 2023.

[17] Pixiv Inc. Pixiv. `https://pixiv.net`. Accessed: 2025-06-09.

[18] Shorten, b., and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data 6*, 1 (2019), 1–48.

[19] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[20] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C. A survey on deep transfer learning, 2018.

[21] Techlabs, M. Use cases of ai-based image recognition. `https://marutitech.com/working-image-recognition/`, 2023.

[22] THBWiki Contributors. Touhou Official Character List. `https://thwiki.cc/%E5%AE%98%E6%96%81%E8%A7%92%E8%89%B2%E5%88%97%E8%A1%A8/%E7%BA%AF%E6%96%87%E5%AD%97%E5%88%97%E8%A1%A8`. Accessed: 2025-06-09.

[23] Touhou Wiki Contributors. Touhou Wiki. `https://en.touhouwiki.net/wiki/Touhou_Wiki`. Accessed: 2025-06-09.

[24] Yang, S., Xiao, W., Zhang, M., Guo, S., Zhao, J., and Shen, F. Image data augmentation for deep learning: A survey, 2022.

[25] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in neural information processing systems* (2014), vol. 27.

[26] Zini, S., Gomez-Villa, A., Buzzelli, M., Twardowski, B., Bagdanov, A. D., and van de Weijer, J. Planckian jitter: countering the color-crippling effects of color jitter on self-supervised training, 2022.