# Distilled Mobile ViT for VizWiz Few-Shot Challenge 2024

Hyunhak Shin, Yongkeun Yun, Dohyung Kim, Jihoon Seo, Kyusam Oh
Eco AI, SK ecoplant
{shin.hyunhak, yongkeun.yun, bryankim82, jihoonseo, q3.oh}@sk.com

## Abstract

*In this paper, we present an efficient lightweight model that leverages knowledge distillation (KD) on mobile vision transformers (ViT) for the VizWiz Few-Shot Challenge 2024. To achieve enhanced performance under constrained computational resources, an optimized training architecture incorporating knowledge distillation is proposed.*

*Our proposed methodology comprises an investigation into adapting vision transformer architectures for mobile deployment, combined with automatic data augmentation techniques and knowledge distillation methods for few-shot learning. Knowledge distillation facilitates transferring knowledge from a larger teacher model to a compact student model, enabling efficient few-shot learning. Our results on the VizWiz Few-Shot Challenge 2024 demonstrate the efficacy of our approach, with the distilled mobile ViT model achieving 71.67% accuracy, ranking among the top 3 submissions in the Few-Shot Object Recognition Task. This substantiates the effectiveness of our methodology in developing efficient and high-performing models for few-shot object recognition under computational constraints.*

## 1. Introduction

Few-shot learning poses significant challenges as it aims to recognize novel objects from only a few labeled samples. Recently, the real-world applicability of few-shot learning has been validated through competitions such as the VizWiz Few-Shot Challenge. The VizWiz Few-Shot Challenge demands accurate models capable of building teachable object recognizers using the ORBIT dataset. The ORBIT dataset [1] provides a highly challenging testbed, capturing diverse objects recorded by individuals with low vision or blindness. Notably, the challenge introduces a new metric requiring models to be computationally lightweight during the personalization steps. Therefore, to avoid accuracy adjustment due to the computational cost of personalization, it is essential to explore efficient lightweight network architectures for personalization.

Our approach focuses on developing an efficient network architecture and learning methodology based on knowledge distillation to improve few-shot learning accuracy. Among the lightweight backbones for image classification, the MobileOne [4] backbone is incorporated due to its proven generalization performance. Additionally, well-known auto-augmentation policies [5-7] are investigated in the training process. Finally, utilizing the DinoV2 model [10] as a teacher model, knowledge is distilled to improve the personalized accuracy for few-shot object recognition.

## 2. Methodology

### 2.1. Network Architecture

To attain high object recognition performance in complex environments with limited computational resources, a vision transformer based mobile network architecture was employed. Following the success of traditional MobileNet, recent works have explored integrating ViT modules into mobile-friendly architectures, such as MobileViT [2], EfficientFormer [3], and MobileOne [4], demonstrating promising results on various vision tasks. Among these architectures, MobileOne-S2 was identified as the most suitable backbone for our few-shot object recognition task. The personalization process of the MobileOne-S2 model was modified to adhere to the VizWiz challenge's criterion of maintaining a computational cost below 1.5T MACs, thereby preventing any accuracy adjustments.

### 2.2. Data Augmentation Strategy

To enhance generalization performance on the VizWiz Few-Shot Challenge, various data augmentation techniques were investigated. However, the vast combinations of augmentations rendered an exhaustive search infeasible during the challenge periods. Consequently, recent research on identifying automatic augmentation policies was consulted. Among the state-of-the-art automatic augmentation methods, Fast AutoAugment [5], RandAugment [6], and Trivial AutoAugment [7] were evaluated. Based on our empirical evaluations, Trivial AutoAugment [7] was found to yield the most robust performance gains for our few-shot object recognition task.

### 2.3. Knowledge Distillation Framework

To boost generalization performance without incurring additional inference costs, a knowledge distillation technique was leveraged. For the teacher model, several large models were evaluated, including ViT [8], CLIP [9], and DinoV2 [10]. Based on our experiments, DinoV2, a large model for vision tasks, was identified as the most suitable teacher for our few-shot object recognition task. However, due to the significant architectural differences and model capacity gap between the large DinoV2 teacher and our compact MobileOne-S2 student model, the benefits of standard knowledge distillation were limited. To mitigate this issue, a recently proposed technique [11] that normalizes the logits during the distillation process was incorporated.

### 2.4. Results Analysis

| Method | MACs | Accuracy (%) | Adjusted Accuracy (%) |
|---|---|---|---|
| EfficienNet-B0 (Baseline) | 0.51T | 67.90% | 67.90% |
| ViT-B-Clip (Baseline) | 5.77T | 74.03% | 64.03% |
| MobileOne-S1 | 1.09T | 68.30% | 68.30% |
| MobileOne-S2 | 1.43T | 69.21% | 69.21% |
| MobileOne-S2 + Trivial Aug. | 1.43T | 69.76% | 69.76% |
| **MobileOne-S2 + Trivial Aug. + KD** | **1.43T** | **71.67%** | **71.67%** |

Table 1: Results in the VizWiz Few-shot Challenge 2024.

According to the results in Table 1, the MobileOne-S2 model achieved an accuracy of 69.21% with a personalization computational cost of 1.43T MACs, which avoided penalties on the averaged accuracy. Additionally, by incorporating the Trivial AutoAugment augmentation technique, the accuracy could be improved of about 0.55% while maintaining the computational cost. Furthermore, the accuracy could also be improved of about 1.89% by employing knowledge distillation approach with normalized logits, utilizing DinoV2 as the teacher model. By combining the data augmentation strategy, the efficient MobileOne-S2 backbone, and the enhanced knowledge distillation framework leveraging DinoV2 as the teacher model, our approach achieved significant improvements in generalization performance for few-shot object recognition. Consequently, an accuracy of 71.67% was achieved while maintaining a low computational load.

## 3. Conclusion

In this work, we proposed a novel approach to develop an efficient and accurate few-shot object recognition model for real-world scenarios with limited computational resources. Our methodology combined three key components: (1) an efficient vision transformer-based mobile architecture, (2) an automatic data augmentation strategy, and (3) an enhanced knowledge distillation framework. By effectively combining these techniques, our method achieved an accuracy of 71.67% on the VizWiz Few-Shot Challenge 2024, while maintaining a low computational cost of 1.43T MACs. In the future, we aim to extend this research by developing practical solutions to effectively detect emerging waste within eco-friendly industries, such as material recovery facilities and electric vehicle battery recycling.

## References


[1] Daniela Massiceti et al. Orbit: A real-world few-shot dataset for teachable object recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[2] Mehta, Sachin, and Mohammad Rastegari., Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021.

[3] Chen, Yinpeng, et al. Mobile-former: Bridging mobilenet and transformer. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[4] Vasu, Pavan Kumar Anasosalu, et al. Mobileone: An improved one millisecond mobile backbone. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[5] Lim, Sungbin, et al. Fast auto-augment. *Advances in Neural Information Processing Systems*, vol 32, 2019.

[6] Cubuk, Ekin D., et al. Randaugment: Practical automated data augmentation with a reduced search space. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020.

[7] Müller, Samuel G., and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. *Proceedings of the IEEE/CVF international conference on computer vision*. 2021.

[8] Dosovitskiy, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Radford, Alec, et al. Learning transferable visual models from natural language supervision. *Proceedings of the IEEE/CVF International conference on machine learning. PMLR*, 2021.

[10] Oquab, Maxime, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[11] Sun, Shangquan, et al. Logit standardization in knowledge distillation. *arXiv preprint arXiv*:2403.01427, 2024.