# A Zero-Shot Classification Method Based on Image Enhancement and Multimodal Model Fusion

Jiamin Cao Lingqi Wang Yujie Shang Lingling Li Fang Liu
Wenping Ma
School of Artificial Intelligence, Xidian University

## Abstract

*We propose a zero-shot image classification method based on data enhancement and multi-model fusion strategy. Since there are no training labels for this dataset, we analyzed the dataset and did low light enhancement and motion blur recovery on the dataset images to help the image feature extractor extract better features. On the processed dataset, we tried various feature extractors based on the timm library, where the tf_efficientnet model achieved a score of total 60.15 in the test-dev channel. In addition, due to the fixed class of the dataset, we tried clip-based multimodal recognition methods, in which the ViT-H-14-378-quickgelu model pre-trained on dfn5b achieves the best result of zero-shot inference among the tested models, with a score of total 61.35 in the test-dev channel. Finally, we asymptotically fuse all models tested, resulting in a final score of 63.9 on test-dev and a score of 65.85 when submitted to the challenge channel.*

## 1. Introduction

Traditional image classification tasks typically require models to be trained on an already labeled training set and then tested and evaluated on a test set. Recently, a more challenging task has emerged, namely Zero-Shot Image Classification, whose goal is to classify images of categories never seen by the model. In this paper, we propose a progressive model fusion strategy approach based on image enhancement and multimodality, which first performs image enhancement on the dataset, and then improves the zero-shot classification effect by fusing the unimodal base model and the multimodal base model. Our contributions are as follows.

(1) We propose an enhancement method for the VizWiz dataset, which experiments have shown to be crucial for improving classification performance.

(2) We introduce an incremental model fusion approach, which allows us to fully utilize different models to improve the accuracy of the prediction results.

## 2. Method

The flow of our approach is shown in Fig.1. First, we perform low-light enhancement and motion blur recovery on the VizWiz image classification dataset[1], which effectively improves the model inference. Second, we extract features both image and image-text modal, and finally, we perform progressive fusion of model predictions, which effectively utilizes all models and improves the final score.

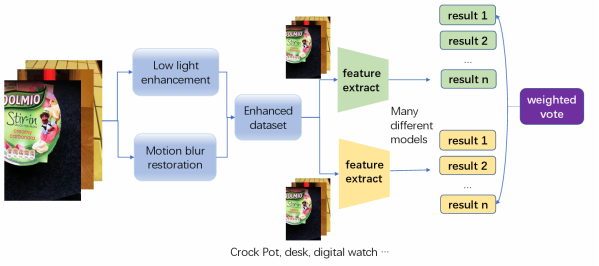


Figure 1. Flowchart of our approach

### 2.1. Image enhanncement

Low light enhancement and motion blur recovery are performed using InstructIR[3] and LAKDNet[6] models respectively. The results are shown in Fig. 2. Since models such as CLIP[5] have text recognition capability, the text of the image after motion recovery is clearer, which facilitates the feature extraction of the CLIP model.

### 2.2. Models

We chose unimodal and multimodal models, where the unimodal model is a model based on the timm library, which achieves a higher accuracy. The tf_efficientnet_l2.ns_jft_in1k_475[7] model we found works better than other models we try, such as volo[8]. Among the multimodal models, we refer to the open-clip-ViT-H-14-

Table 1. Comparison of results before and after using image enhancement

| data | modal | model | total score |
|---|---|---|---|
| origin | image | tf_efficientnet_l2.ns_jft_in1k_475 | 59.55 |
| enhanced | image | tf_efficientnet_l2.ns_jft_in1k_475 | 60.15 |
| origin | image+text | open-clip-ViT-H-14-378-quickgelu | 60.4 |
| motion enhanced | image+text | open-clip-ViT-H-14-378-quickgelu | 60.65 |
| enhanced | image+text | open-clip-ViT-H-14-378-quickgelu | 61.35 |
| origin | image | eva_giant_patch14_560.m30m_ft_in22k_in1k | 57.15 |
| enhanced | image | eva_giant_patch14_560.m30m_ft_in22k_in1k | 58.5 |

(a) before lol (b) after lol (c) before mbr (d) after mbr

Figure 2. Our image enhancement examples, lol means low light enhancement, mbr means motion blur recovery.

Table 2. Model fusion results

| model | score |
|---|---|
| ViT-H-14-378-quickgelu-enhanced | 61.35 |
| tf_efficientnet_l2.ns_jft_in1k_475 | 60.15 |
| open-clip-ViT-H-14-378-quickgelu-motion | 60.65 |
| weight fusion | 63.9 |

378-quickgelu[4] model with the strongest zero-shot normalization ability among the open_clip[2] models.

### 2.3. Progressive model fusion

We vote on different models with similar scores and then do a weighted vote on the fusion results to get the final result.

## 3. Experiments

### 3.1. Effectiveness of data augmentation

For the same model, using our data enhancement improves total score by 0.6%-1.35%. All scores in Table 1 are total scores over test-dev.

### 3.2. Effectiveness of model fusion

In order to fully utilize the models we tried, we voted and weighted the fusion of models with different score bands. All scores in Table 2 are total scores over test-dev.We submit the weight fusion result to challenge submit channel, and achieve 65.85 total score.

## 4. Conclusion

We performed effective data augmentation based on the VizWiz dataset, using multiple models for inference based on the modality of images and image-text pairs, and progressive fusion of the results to fully utilize multiple models and effectively improve the final prediction score.

## References


[1] Reza Akbarian Bafghi and Danna Gurari. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16261–16270, 2023. 1

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 2

[3] Marcos V Conde, Gregor Geigle, and Radu Timofte. High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2024. 1

[4] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023. 2

[5] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[6] Lingyan Ruan, Mojtaba Bemana, Hans-peter Seidel, Karol Myszkowski, and Bin Chen. Revisiting image deblurring with an efficient convnet. *arXiv preprint arXiv:2302.02234*, 2023. 1

[7] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 1

[8] Li Yuan, Qibin Hou, Zihang Jiang, Jiashi Feng, and Shuicheng Yan. Volo: Vision outlooker for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45 (5):6575–6586, 2022. 1