

Supplementary Materials: WaveDN: A Wavelet-based Training-free Zero-shot Enhancement for Vision-Language Models

Anonymous Authors

A DATASETS DETAILS

The sizes of the test set, the textual templates employed, and the configurations of the evaluation metrics pertinent to the dataset utilized in our study are systematically delineated in Table. 1. To bolster the reproducibility and precision of the experimental outcomes, this section will elucidate the specifics concerning the utilization of the dataset.

A.1 Image Datasets

In this paper, the image datasets utilized are widely recognized in the field of computer vision and serve as established benchmarks for image classification models. The CLIP text encoder relies on text templates for guidance. During our experiments, ImageNet[1], CIFAR-100[2], Stanford Cars[3], SUN397[4], and Food-101[5] all employ the standard text template "a photo of a [Label]," which is a widely adopted format. On the other hand, DTD[6], Oxford Pets[7], Oxford Flowers 102[8], and EuroSAT[9] employ customized text templates tailored to the unique characteristics of each dataset, mirroring the approach in CALIP[10]. While Oxford Pets and Oxford Flowers 102 are assessed using "Mean per class accuracy," the remaining datasets are evaluated using "Accuracy," aligning with the methodology employed in CLIP[11].

The MSCOCO[12] dataset is a versatile image dataset commonly employed for tasks like object detection, segmentation, and image captioning. In our retrieval task, we utilized the image captioning annotations based on the Karpathy split[13], with a test set size of

5000 images. On the other hand, the Flickr30k[14] dataset serves as a benchmark dataset comprising numerous images paired with corresponding descriptions. We adopted a data split identical to that of research [15] and research[16], resulting in a test set size of 1000 images.

A.2 Audio Datasets

The UrbanSound8K[17] dataset is a compilation of audio data consisting of 10 distinct classes and a total of 8732 audio tracks, each track not exceeding 4 seconds in duration. The prevalent audio categories encompass commonplace sounds like "air conditioner," "car horn," and "children playing" typically encountered in daily life. The ESC-50[18] dataset comprises 50 audio categories and a total of 2000 tracks, each track spanning a duration of 5 seconds. The primary audio categories within this dataset encompass five major groups: animal sounds, natural and water sounds, non-speech human sounds, interior sounds, and exterior sounds. We aligned our partitioning of the UrbanSound8K dataset with research[17] and the ESC-50 dataset with research[18]. In the data preprocessing stage, we sampled all audio files at a rate of 44100Hz and transformed the sampled audio vectors into 1D tensors for input into the audio encoder.

B CLASS-LEVEL ACCURACY ANALYSIS

Comparison of class-level results on all image datasets on Vit-B32 using two methods, DN and WaveDN, is shown in Figure. 1. The

Table 1: Dataset details

Dataset	Test split size	Text template	Evaluation metric
ImageNet	50000	"a photo of a [Lable]"	Accuracy
Cifar100	10000	"a photo of a [Lable]"	Accuracy
Standford Cars	8041	"a photo of a [Lable]"	Accuracy
sun397	19850	"a photo of a [Lable]"	Accuracy
DTD	1880	"[Lable] texture"	Accuracy
Oxford Pets	3669	"a photo of a [Lable], a type of pet"	Mean per class accuracy
Food-101	25250	"a photo of a [Lable]"	Accuracy
Oxford Flowers 102	6149	"a photo of a [Lable], a type of flower"	Mean per class accuracy
EuroSAT	5000	"a centered satellite photo of [Lable]"	Accuracy
FGVC Aircraft	3333	"a photo of a [Lable], a type of aircraft"	Accuracy
ESC-50	2000	"[Lable]"	Accuracy
UrbanSound8k	8732	"[Lable]"	Accuracy
ESC-50(retrival)	2000	"[Lable]"	Recall
UrbanSound8k(retrival)	8732	"[Lable]"	Recall
MSCOCO	5000	"A photo depicts [Discription]"	Recall
Flicker30k	1000	"A photo depicts [Discription]"	Recall

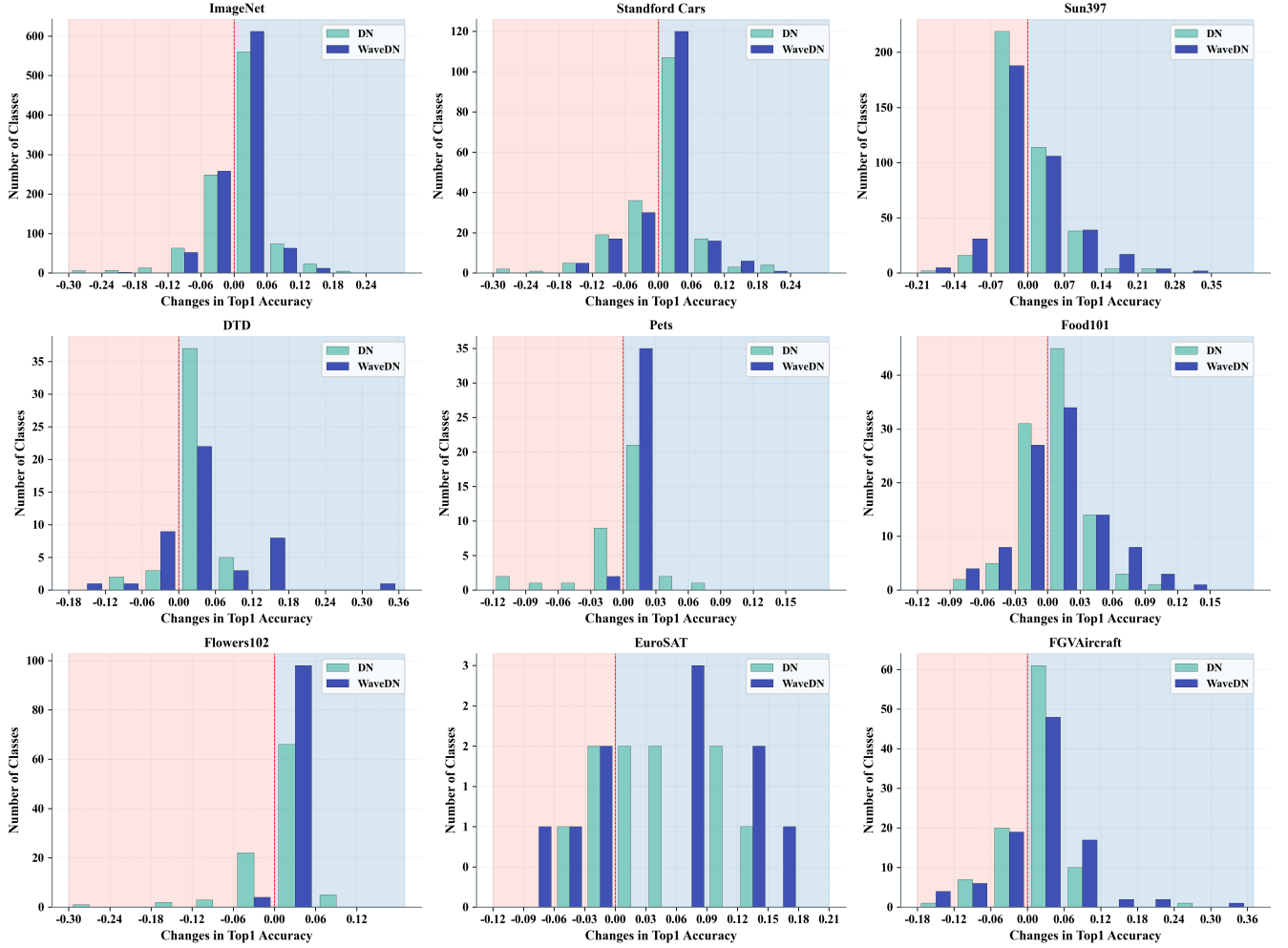


Figure 1: Comparison of Class-Level Top-1 Accuracy Changes Using ViT-B32.

class-level comparison results on Cifar100 are reported in the main text, here we present the class-level accuracy variation results for the remaining image datasets. The bar graphs with light red background indicate the number of classes where the recognition accuracy decreases after applying a certain method. Conversely, the bars with light blue background represent the number of classes where the recognition accuracy increases. Across numerous datasets, it is evident that WaveDN significantly reduces the number of classes affected by negativity and has a more positive impact on a greater number of classes. These experimental results demonstrate that compared to DN, WaveDN disrupts the original data representation in embedding less and can achieve alignment with InfoNCE loss during the testing phase.

In the experiments with EuroSAT and Food101 datasets, WaveDN has a higher negative impact compared to DN. However, the reduction in recognition accuracy by WaveDN is at a relatively low level, not exceeding 10%. Moreover, for these datasets, WaveDN exerts a stronger positive impact, with dataset-level accuracy higher than DN. Although WaveDN may also negatively affect the recognition

accuracy of certain categories, the magnitude of the negative impact is much smaller than DN. From the graphs, it is evident that DN can decrease the recognition accuracy of some categories by over 30%, which severely undermines the generalization ability of VLM itself.

In the experimental results of FGVAircraft, WaveDN has some additional negative impacts compared to DN (eg. -0.18 to -0.12). We believe this is related to the distribution characteristics of the FGVAircraft dataset itself, where the representations of aircraft images in the dataset are highly similar, leading to very close similarities between the embeddings output by the CLIP image encoder. This similarity makes CLIP less adaptable to this dataset. Due to the model's poor understanding of the data within the dataset, WaveDN cannot perform well, resulting in additional negative impacts to some classes. However, at the dataset level, WaveDN still has a good positive effect on recognition accuracy, improving recognition accuracy more than DN.

REFERENCES

- [1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Jun 2009.
- [2] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.
- [3] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, Dec 2013.
- [4] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun 2010.
- [5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. *Food-101 – Mining Discriminative Components with Random Forests*, page 446–461. Jan 2014.
- [6] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.
- [7] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- [8] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.
- [9] 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*, page 2217–2226, Jul 2019.
- [10] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 746–754, 2023.
- [11] Alec Radford, JongWook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Askell Amanda, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *Cornell University - arXiv, Cornell University - arXiv*, Feb 2021.
- [12] 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*, page 740–755. Jan 2014.
- [13] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.
- [14] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, page 74–93, May 2017.
- [15] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.
- [16] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [17] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.
- [18] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.