Supplementary Materials for "Dynamic MLP for Fine-Grained Image Classification"

Lingfeng Yang[1†], Xiang Li[1*], Renjie Song[2], Borui Zhao[2], Juntian Tao[1], Shihao Zhou[2], Jiajun Liang[2], Jian Yang[1*]
[1]Nanjing University of Science and Technology, [2]Megvii Technology
{yanglfnjust, xiang.li.implus, taojuntian, csjyang}@njust.edu.cn
{songrenjie, zhoushihao, liangjiajun}@megvii.com, zhaoborui.gm@gmail.com

A. Datasets

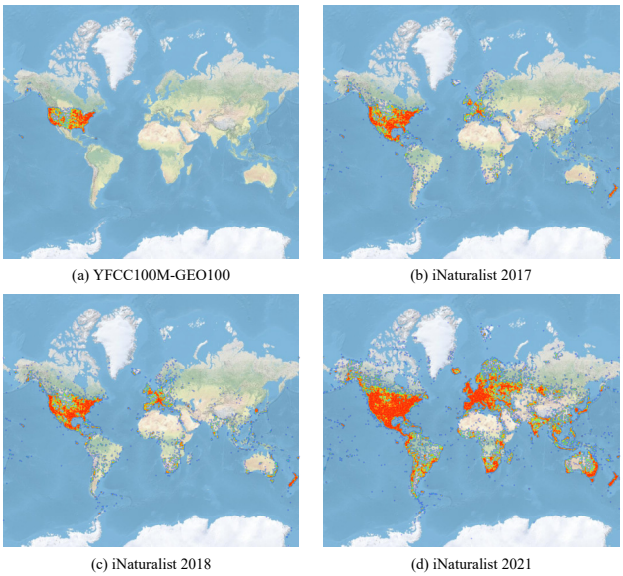
In this section, we introduce the fine-grained datasets with geolocation and dates in detail. Notably, we visualize the data distribution as a heatmap on the iNaturalist and YFCC100M-GEO100 datasets to depict the geographical information (Fig. S1).

A.1. iNaturalist

We perform most of the experiments on the iNaturalist 2017, 2018, and 2021 [17, 18] datasets with geographical and temporal information from the fine-grained image classification challenge at FGVC (fine-grained visual categorization). The iNaturalist datasets contain various species photographed by the public and then identified and annotated by experts at [5]. The iNaturalist 2017 dataset has 579,184 training data and 95,986 validation data with 5,089 categories, while the iNaturalist 2018 dataset has 437,513 and 24,426 images for training and validation within 8,142 labels. The latest iNaturalist 2021 dataset has 2,686,843 training data points in one of its subsets, i.e., 500,000 images marked as the iNaturalist 2021 mini dataset, and 100,000 images for validation with 10,000 categories.

A.2. YFCC100M

Authorized by Flickr, the YFCC100M [15] dataset consists of 100 million images where approximately 49 million images are annotated with geographical information, e.g., latitude and longitude. Several works [1,11,14] have chosen more qualified and illustrative images to form YFCC100M subsets. [14] selects the top 100 classes according to location sensitivity and forms the YFCC100M-GEO100 dataset.



(a) YFCC100M-GEO100 (b) iNaturalist 2017 (c) iNaturalist 2018 (d) iNaturalist 2021

Figure S1. The geographical distribution heatmap of the iNaturalist and YFCC100M-GEO100 datasets.

It contains 88,986 images and is then split into 86,986 training data and 2,000 validation data, i.e., 20 images per class for 100 categories in our experiments. In GeoNet [1], labels of YFCC100M that are related to the corresponding species within the iNaturalist 2017 dataset are collected to form the YFCC100M-Geolocated-iNat2017species dataset. There are 36,143 images in total, with 4,472 categories, where each label has at least one instance. Notably, we use the YFCC100M-GEO100 dataset for our experiments.

B. More Training/Test Details

B.1. Training

During training, we apply a random crop of 224×224 pixels, a random horizon flip [13], Mixup [20], and label smoothing to the inputs as data augmentations. All CNN

*Corresponding author. †Works is done as interns in Megvii Research. Lingfeng Yang, Xiang Li, Juntian Tao, and Jian Yang are from PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology.

Backbone	R_{train}	R_{test}	Ten-crop	Acc (%)
SK-101	224	224		91.397
SK-101	224	384		92.028
SK-101	224	256, 288, 320, 352, 384	✓	92.609
SK-101	384	448, 480, 512, 544, 576	✓	93.101
SK-101	448	512, 544, 576, 608, 640	✓	93.712
BoT-152	224	256, 288, 320, 352, 384	✓	91.930
BoT-152	384	448, 480, 512, 544, 576	✓	92.379
Swin-Large	224	224	✓	92.357
PVT-Large	224	256, 288, 320, 352, 384	✓	92.879
Ensemble				94.750

Table S1. Our model results in FGVC8 [2] on the iNaturalist 2021 dataset. "R_{train}" and "R_{test}" denote the training and testing resolution, respectively. **SK**: SK-Res2Net [3, 7]. **BoT**: BoTNet [12]. **Swin**: Swin Transformer [8]. **PVT**: [19]. **Ensemble**: The ensemble result including models with an accuracy higher than 91.00%.

backbone networks are trained using SGD with a momentum of 0.9, a weight decay of 1×10^{-4}, and 8 GPUs with a mini-batch size of 64 on each to optimize models. The learning rate is set to 4×10^{-2} with a linear warmup [4] for two epochs and a cosine decay schedule [9]. To optimize models for Transformer backbones, we use AdamW [10] with a momentum of 0.9, a weight decay of 5×10^{-2}, and a mini-batch size of 32, with an initial learning rate of 2×10^{-4}. The augmentation follows the descriptions in PVT [19]. We train 60 epochs for the iNaturalist 2021 full dataset and 90 epochs for other datasets.

B.2. Inference

During inference, a center crop is applied to the image as data augmentation. The testing resolution is aligned with the training phase. For all multimodal methods, both images and additional information are utilized for evaluation.

C. More Details for FGVC8 Competition

During the FGVC8 competition [2], we leverage several powerful CNN [3, 7, 12] and Transformer [8, 19] models as the backbones of our framework. The training and testing settings are illustrated in Sec.B. We also evaluate models using the multi-scale strategy, where R_{test} has multiple values. As indicated by FixRes [16], a higher resolution in testing than in the training phase would improve the accuracy. We only apply Ten-crop [6] as a post-processing strategy to the Swin Transformer [8] since it requires fixed model input.

D. Re-implementation of former works

We follow the official repository[1] to reproduce PriorsNet [11]. This method first trains a separate CNN model based on solely images. Then it additionally trains a location encoder, a fully-connected neural network consisting of an input layer, followed by multiple residual layers, and a final output embedding layer. Refer to detailed training settings in PriorsNet [11] for more information. Specifically, the image-only model is irrelevant to the core implementation of PriorsNet, i.e., the location encoder, which only takes as input the locations and dates. Further, the training pipeline is available in the codebase. In our experiments, the MLP backbone of the multimodal path in our dynamic MLP is aligned with the location encoder in PriorsNet for a fair comparison. Next, a unified image classifier based on the popular backbone is well-trained on image data. In the case where optimal training settings are different for multiple datasets, we conduct a series of ablation studies on the hyperparameters of the location encoder in PriorsNet. Officially, the location encoder is trained with a batch size of 1024 and a learning rate of 5×10^{-4} on one GPU. A class-balanced (CB) sampler is used to counteract the imbalanced nature of long-tailed datasets where the image number per class is set to 100 by default. Considering the iNaturalist 2021 dataset is class-balanced, we also attempt to use an instance-balanced (IB) sampler for the data loader. The final results of PriorsNet [11] and our dynamic MLP are reported in Table S2.

Method	Batch	lr	Sampler	N / C	Acc (%)
PriorsNet [11]	1024	5×10^{-4}	CB	100	80.246
	2048	1×10^{-3}	CB	100	80.275
	4096	2×10^{-3}	CB	100	80.302
	1024	5×10^{-4}	CB	50	80.330
	2048	1×10^{-3}	CB	50	80.289
	4096	2×10^{-3}	CB	50	80.302
	1024	5×10^{-4}	IB	–	80.294
	2048	1×10^{-3}	IB	–	80.271
	2048	2×10^{-3}	IB	–	80.288
PriorsNet [11]*	1024	5×10^{-4}	CB	50	83.600
Dynamic MLP **(ours)**					**84.694**

Table S2. Comparisons of different batch sizes, learning rates, and samplers on the iNaturalist 2021 mini dataset. **CB**: Class-balanced sampler. **IB**: Instance-balanced sampler. "N / C" is the image number per class for the class-balanced sampler. *The location prior of the full dataset is used to promote the prediction for the mini dataset (not allowed essentially).

Next, we elaborate on our re-implementation of the GeoNet [1], with reference to the official repository[2] and the baseline[3] they have compared. Specifically, we refer to the post-processing strategy that achieves the highest results reported in the paper. The post-processing models are trained with a learning rate of 0.02 without decay, based on a well-pretrained image-only model. Since the fine-tune epoch is not specified, we train the network until no further performance improvement can be obtained, which is 90 epochs specifically. As shown in Fig. S2, the best performances are achieved before 60 epochs for iNaturalist 2017, 2018, and 2021 [17, 18] generally.

[1] https://github.com/macaodha/geo_prior

[2] https://github.com/visipedia/fg_geo

[3] https://github.com/richardaecn/cvpr18-inaturalist-transfer

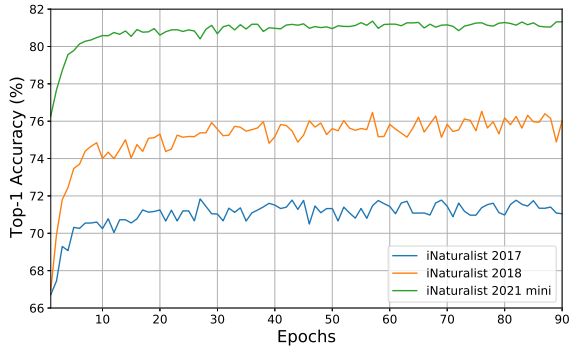


Figure S2. The top-1 accuracy of GeoNet [1] on iNaturalist 2017, 2018, and 2021 [17, 18]. Note that the initial accuracy is inherited from the image-only pretrained models.

ResNet-50	#Params	Flops	YFCC	iNat18
Attenion (Q: img, K,V: geo-tem)	48.1 M	4.1 G	52.150	77.282
Attenion (Q: geo-tem, K,V: img)	48.1 M	4.1 G	52.325	78.093
Attenion (Q,K,V: concat both)	49.6 M	4.1 G	52.550	77.340
Dynamic MLP **(ours)**	47.4 M	4.1 G	**53.200**	**78.220**

Table S3. Comparisons to the scaled dot-product attention variants on YFCC and iNaturalist 2018 under ResNet-50 backbone.

E. More Experimental Results

E.1. Comparison with Attention Module

Taking image and geo-temporal feature vectors as inputs, we implement the scaled dot-product attention whilst keeping all other MLP structures the same with dynamic MLP. In the first row of Table S3, we set the image feature as Q, geo-temporal feature as K and V, and exchange their positions in the second row. In the third row, we use their concatenation as Q,K, and V. It is observed that the proposed dynamic MLP is more accurate.

E.2. More Results on Fusion Strategies

Beyond Table 6, we show more experimental results on YFCC and iNaturalist 2018 under ResNet-50 backbone in Table S4, where our method achieves all the best. We also conduct an extra ablation study to compare all methods under the exact same MLP structures in Table S5, where we only change the mat-multiply to other operations. It is observed that dynamic MLP still keeps its superiority.

E.3. Individual Benefits of Additional Information

Table S6 shows the individual gain of geographical/temporal information on the iNaturalist 2018 and 2021 mini datasets under SK-Res2Net-101 backbone, where the geographical information makes a major contribution to the overall improvements.

Backbone	Method	#Params	Flops	YFCC	iNat18
R-50	Concatenation*	47.4 M	4.1 G	51.050	76.537
	Addition*	47.4 M	4.1 G	50.950	77.139
	Multiplication*	47.4 M	4.1 G	52.050	76.394
	Dynamic MLP **(ours)**	47.4 M	4.1 G	**53.200**	**78.220**
SK-101	Concatenation*	70.0 M	8.9 G	55.100	81.892
	Addition*	70.0 M	8.9 G	54.650	82.334
	Multiplication*	70.0 M	8.9 G	54.450	81.151
	Dynamic MLP **(ours)**	70.0 M	8.9 G	**56.800**	**83.673**

Table S4. Comparisons to other fusion strategies under the same complexity. **R-50**:ResNet-50. **SK-101**: SK-Res2Net-101.

ResNet-50	#Params	Flops	YFCC	iNat18
Concatenation†	47.4 M	4.1 G	52.100	77.585
Addition†	47.4 M	4.1 G	52.750	77.569
Multiplication†	47.4 M	4.1 G	52.500	77.528
Dynamic MLP **(ours)**	47.4 M	4.1 G	**53.200**	**78.220**

Table S5. Comparisons to other fusion strategies under the same MLP structures on iNaturalist 2018 under ResNet-50 backbone.

Geo.	Tem.	iNat18	iNat21
–	–	74.150	76.102
✓		83.624 (+9.47)	84.310 (+8.21)
	✓	75.493 (+1.34)	76.821 (+0.72)
✓	✓	**83.673** (+9.52)	**84.694** (+8.59)

Table S6. Ablation study on individual gain of geographical/temporal information on iNaturalist 2018 and 2021 mini under SK-Res2Net-101 backbone. **Geo.**: Geographical information. **Tem.**: Temporal information.

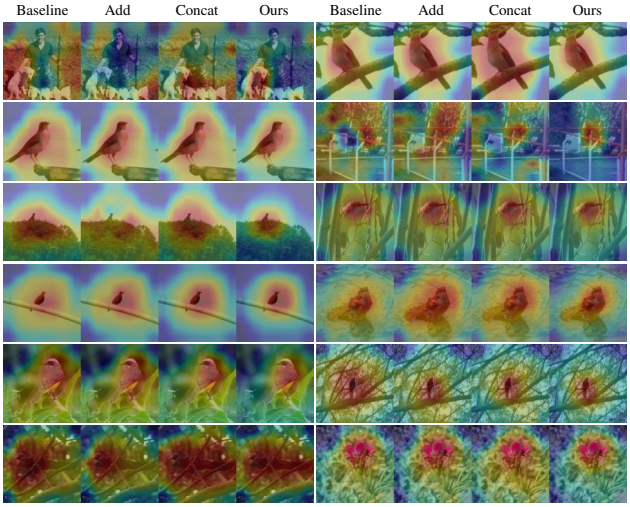


Figure S3. The visualization of the activation heatmap on the images under the ground-truth label. Our methods can precisely locate the semantically meaningful regions, especially when the image is more complicated.

F. More Analysis for Dynamic MLP

F.1. Visualization of the Activation Map

In Fig. S3, we visualize the activation map under the ground-truth label between the previous works and dynamic

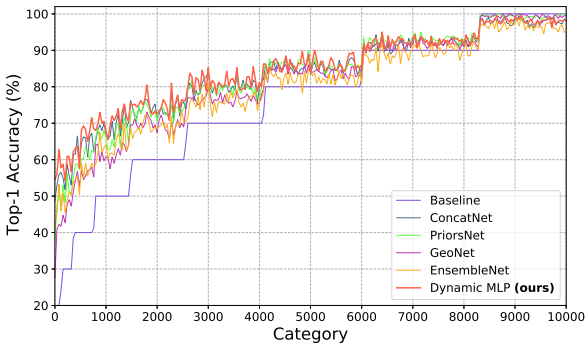


Figure S4. Comparisons of the top-1 accuracy per category on the iNaturalist 2021 dataset among various methods. Dynamic MLP achieves superior performance in the majority of categories.

MLP using CAM [21]. It demonstrates that dynamic MLP can learn more precise, concentrated, and reasonable activation maps on images, which potentially benefits the image representation under the guidance of extra information.

F.2. Accuracy per Categories

Fig. S4 indicates the top-1 accuracy for each category under various methods on the iNaturalist 2021 dataset. For better visualization, we rank the categories by the accuracy of the baseline, from low to high, and average the accuracy of 100 adjacent categories. Since each label in the iNaturalist 2021 dataset contains 10 images, Fig. S4 presents a stepped curve. Our dynamic MLP boosts the performance for most of the categories, which is typically obvious on the hard labels (low accuracy in the baseline). Notably, all methods that utilize the additional information produce a few wrong predictions on instances that are correctly classified by image-only models, which indicates our method can still be improved further.

References

[1] Grace Chu, Brian Potetz, Weijun Wang, Andrew Howard, Yang Song, Fernando Brucher, Thomas Leung, and Hartwig Adam. Geo-aware networks for fine-grained recognition. In *ICCV*, 2019. 1, 2, 3

[2] FGVC8. https://www.kaggle.com/c/inaturalist-2021, 2021. 2

[3] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 2019. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[5] iNaturalist. http://www.inaturalist.org, 2008. 1

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 2

[7] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, 2019. 2

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2

[9] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 2

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[11] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *ICCV*, 2019. 1, 2

[12] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *CVPR*, 2021. 2

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 1

[14] Kevin Tang, Manohar Paluri, Li Fei-Fei, Rob Fergus, and Lubomir Bourdev. Improving image classification with location context. In *ICCV*, 2015. 1

[15] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 2016. 1

[16] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. Fixing the train-test resolution discrepancy. In *NeurIPS*, 2019. 2

[17] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. *arXiv preprint arXiv:2103.16483*, 2021. 1, 2, 3

[18] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 2, 3

[19] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2

[20] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1

[21] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 4