Supplementary materials: LLM-assisted Entropy-based Adaptive Distillation for Self-Supervised Fine-Grained Visual Representation Learning

Jianfeng Dong[1,5] Danfeng Luo[1] Daizong Liu[2†] Jie Sun[1,5†]
Xiaoye Qu[3] Xun Yang[4] Dongsheng Liu[1,5] Xun Wang[1,5]
[1]Zhejiang Gongshang University, [2]Peking University
[3]Huazhong University of Science and Technology, [4]University of Science and Technology of China
[5]Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology
https://github.com/HuiGuanLab/LEAD

1. Different Distillation Methods

Figure 1 summarizes the comparison of our LLM-assisted distillation with two widely used distillation methods: feature-based distillation [4], relation-based distillation [7, 9]. The effectiveness of our LLM-assisted logits-based distillation lies in its ability to transfer the teacher model's knowledge to the student model by constructing a category-wise probability distribution grounded in LLM-generated text descriptions of categories. This approach allows the model to better align the teacher's knowledge with the specific requirements of fine-grained downstream tasks. We attribute the superior performance of our method to the use of category-wise probability distributions, which not only preserve the discriminative power of the teacher model but also adapt its knowledge to fine-grained scenarios with the assistance of LLM-generated descriptive guidance.

2. Influence of the Number of Images

Based on the experimental results in Figure 2, which examine the impact of the number of training images on model performance, we observe that as the number of training images increases, both our method, LEAD, and the baseline, MoCo, exhibit gradual improvements in Top-1 image classification accuracy. However, LEAD consistently outperforms MoCo across all data scales. This demonstrates that LEAD is capable of effectively extracting fine-grained features even under limited data conditions, fully utilizing the information from each image, thereby exhibiting superior generalization ability and robustness.In contrast, MoCo's performance is constrained in small-sample scenarios, where it struggles to capture subtle differences between fine-grained categories. Furthermore, when the number of training images reaches a larger scale (e.g., 5000 images or more), the performance improvement of LEAD becomes even more pronounced. This highlights LEAD's advantage in leveraging the multimodal capabilities of LLM and CLIP, enabling it to further explore the potential of fine-grained features in large-scale data settings. These results validate the effectiveness and superiority of our proposed method in fine-grained visual representation learning tasks.

Table 1. Performance of using different LLMs.

LLM	Classification		Image Retrieval		
	Top-1	Top-5	R-1	R-5	mAP
DeepSeek	**78.79**	**95.46**	**68.29**	**89.32**	**44.63**
GPT-3(Ours)	78.44	95.30	68.17	88.37	44.43

3. Influence of Using Different LLMs

We evaluated the performance of our method using different LLMs to assess its sensitivity to the choice of LLMs, as shown in Table 1. Specifically, when using DeepSeek, our method achieved a performance of 78.79%, which is comparable to the 78.44% achieved with GPT-3. This indicates that the choice of LLM has only a minimal impact on the overall performance of our approach. These results suggest that our method is robust and not heavily dependent on a specific LLM for generating descriptions, making it adaptable to a wide range of language models without significant performance degradation. This flexibility further highlights the generalizability and practicality of our proposed framework.

†Corresponding authors: Daizong Liu and Jie Sun.

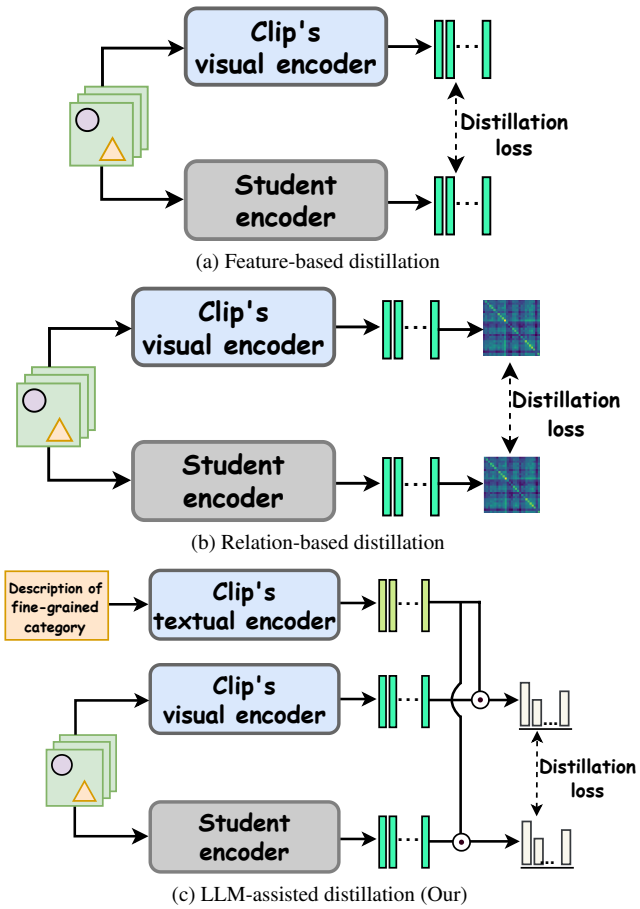


Figure 1. Schematic comparison of (a) traditional feature-based distillation, (b) traditional relation-based distillation, and (c) our proposed LLM-assisted distillation. While traditional methods rely solely on data relationships for knowledge transfer, our approach integrates both data and LLM-generated knowledge, enhancing fine-grained learning.

4. Influence of the Number of Categories

In this experiment, we investigate how the number of categories affects the results. As shown in Figure 3, increasing the number of categories to 250 on the Cub dataset (above the actual 200 categories) leads to a marginal performance improvement, but further increases result in slight performance degradation. We speculate that, as the number of fine-grained categories per class is limited, excessive categories lead to meaningless distinctions and semantic confusion.

5. Influence of Text Length

Table 2 presents the impact of varying text lengths on the performance of our model. The results show that when the text length is relatively short, the model experiences a slight performance drop; however, the decline remains minimal.

Table 2. Performance of different text lengths.

Text Lengths	Classification		Image Retrieval		
	Top-1	Top-5	R-1	R-5	mAP
10	77.15	94.41	66.54	86.78	43.09
20	77.81	95.32	67.63	88.11	44.07
30	77.63	95.12	67.29	87.44	44.00
40	77.91	95.15	67.69	88.37	44.11
50	78.22	95.29	68.14	88.40	**44.51**
60	**78.80**	**95.50**	67.97	**88.49**	44.44
70(Ours)	78.44	95.30	**68.17**	88.37	44.43

Table 3. Performance comparison of CLIP with different backbones on Cub.

teacher's backbones	Classification		Retrieval		
	Top-1	Top-5	R-1	R-5	mAP
CLIP-L/14	**79.22**	**95.62**	**72.06**	**90.25**	**49.27**
CLIP-B/16	78.44	95.30	68.17	88.37	44.43
CLIP-B/32	74.94	94.06	60.79	84.57	34.21

Table 4. Performance of varying word removal ratio on Cub.

Removal Ratio	Classification		Image Retrieval		
	Top-1	Top-5	R-1	R-5	mAP
0%(Ours)	**78.44**	**95.30**	**68.17**	**88.37**	**44.43**
20%	77.84	95.17	65.08	87.12	39.23
40%	76.42	94.81	64.53	86.85	37.99
60%	76.04	94.51	61.58	84.88	32.97
80%	74.54	94.08	57.90	82.55	30.70

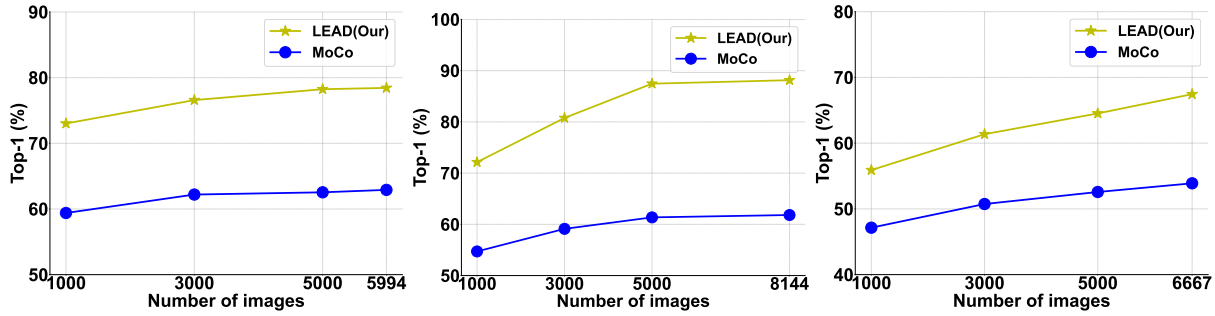
We hypothesize that even with shorter text, LLMs are capable of effectively capturing and describing fine-grained key features, ensuring robust performance. On the other hand, as the text length increases, much of the additional content tends to be redundant or irrelevant, contributing little to further improving the model's performance. These findings suggest that our approach is both efficient and resilient to variations in text length.

6. Influence of Errors in CLIP and LLM

In this experiment, we investigate how errors in CLIP and LLM affects the results. For CLIP, we try different variants of decreasing abilities with more errors (ViT-L/14, ViT-B/16, ViT-B/32) in Table 3. For LLM, we simulate semantic errors by randomly removing words from descriptions. As shown in Table 4, the performance slightly decreases with more word removal. These results demonstrate that errors in CLIP and LLM lead to minor performance degradation.

7. Comparison to IBOT and SimCLR

In order to further verify the effectiveness of our proposed method, we compare two popular self-supervised methods,



(a) CUB (b) Car (c) Aircraft

Figure 2. Performance curve for fine-grained image classification with varying sample sizes. Our method consistently outperforms MoCo across all image quantities, demonstrating its robustness and effectiveness.

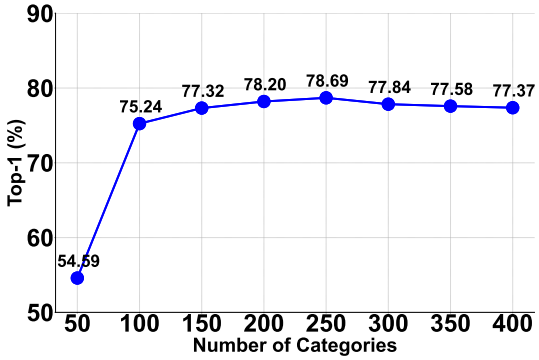


Figure 3. Using varying numbers of categories and text lengths.

Table 5. Performance comparison to IBOT and SimCLR on Cub.

Method	Classification		Retrieval		
	Top-1	Top-5	R-1	R-5	mAP
SimCLR	37.28	66.17	16.43	36.24	4.78
IBOT	73.66	92.37	61.53	83.93	34.89
Ours	**78.44**	**95.30**	**68.17**	**88.37**	**44.43**

i.e. BOT and SimCLR. As shown in Table 5, our proposed method still performs better.

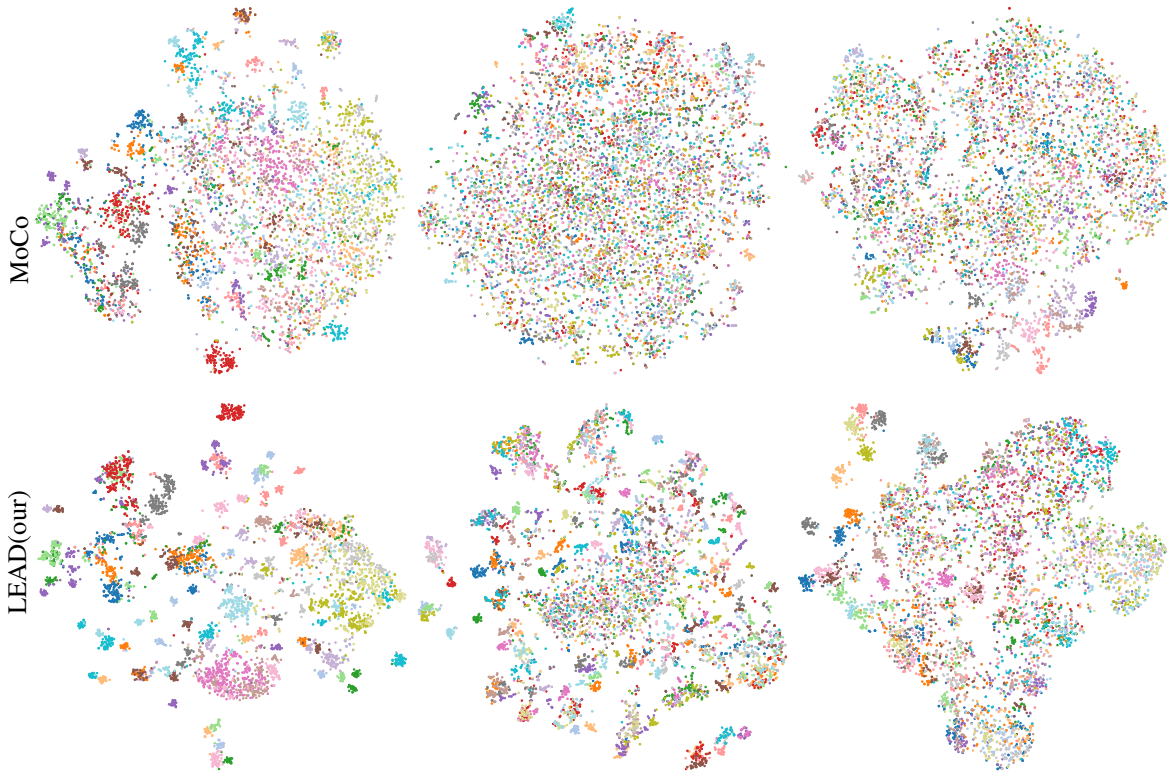
8. More Visualization Results

T-SNE Visualization. Figure 4 presents the t-SNE [10] visualizations of feature representations generated by MoCo and our proposed method on three fine-grained datasets: CUB, Cars, and Aircraft. The visualizations clearly demonstrate that our method achieves superior clustering performance, with samples from the same category tightly grouped and distinct boundaries formed between different categories. In contrast, the feature representations produced by MoCo exhibit significant overlap and poorly defined category boundaries, with these shortcomings being particularly pronounced on the CUB dataset. It is important to note that MoCo is equivalent to our method without the incorporation of knowledge distillation. Through the clearly improved clustering results, it is evident that the integration of knowledge distillation and the adaptive module significantly enhances the separability of features. This improvement is attributed to the knowledge distillation approach, which effectively leverages the semantic information from the CLIP teacher model and the contextual textual knowledge provided by large language models, thereby strengthening the discriminative capability of features across different categories.The experimental results further validate the effectiveness of entropy-based adaptive distillation. This technique not only enhances the quality of feature representations but also significantly improves the model's adaptability and performance in fine-grained tasks.

Visualization of more classification results. Figure 5 demonstrates some representative classification examples. In the first column, the teacher model correctly classifies the sample with low information entropy. Here, the adaptive module assigns greater weight to the KD branch, allowing the student model to effectively learn accurate knowledge from the teacher model. In the second column, the teacher model produces an incorrect prediction accompanied by high information entropy. To mitigate the impact of erroneous knowledge, the adaptive module dynamically shifts focus toward the CL branch, enabling the student model to make the correct classification despite the teacher model's failure. These examples demonstrate the adaptive module's ability to balance the contributions of the KD and CL branches, optimizing knowledge transfer and improving classification performance in varied scenarios. In the third section, we also present two challenging examples where the distinguishing features are not obvious, causing our model to struggle with handling them.

9. Descriptions Generated by LLM

Table 6 presents examples of fine-grained labels and their corresponding textual descriptions generated by LLM. As



(a) Cub (b) Car (c) Aircraft

Figure 4. t-SNE visualization of MoCo and our method on three datasets. Dots of the same color represent images sharing the same class label.Our method demonstrates significantly better clustering performance, with samples of the same category closely grouped and clear boundaries formed between different categories. In contrast, the feature representations of MoCo v2 exhibit substantial overlap and poorly defined category boundaries.

can be observed from the table, the textual descriptions generated by the LLM provide detailed and nuanced descriptions of the visual characteristics specific to each category. These descriptions effectively capture fine-grained semantic and visual features, enabling the pre-trained CLIP model to better leverage its potential in representing fine-grained distinctions. By incorporating these detailed and context-rich textual descriptions, the framework facilitates the transfer of more comprehensive and structured knowledge to the student model, ultimately improving its ability to understand and represent fine-grained features. This approach highlights the significant role of LLM-generated descriptions in bridging the semantic gap and enhancing the performance of downstream fine-grained tasks.

10. Evaluation Protocols

Following the previous works [1, 8, 11], we evaluate the effectiveness of the proposed method in two downstream tasks: fine-grained image classification and fine-grained image retrieval. In these two downstream tasks, each model should be first trained in a self-supervised manner without using any labeled data, and then used for further evaluation.

In the image classification evaluation, the parameters of the model trained by the self-supervised representation learning method are fixed and a linear classifier is attached to it. The linear classifier is trained to perform classification and its classification performance reflects the quality of the self-learned representation. We utilize the Top-1 and Top-5 accuracy as the classification performance.

Image retrieval (also equivalent to the nearest neighbor classification) aims to search for neighbors close to the query image in the latent feature space without adjusting any model parameters. We use Rank-1, Rank-5, and mean Average Precision (mAP) as the performance metrics. For ease of reference, we simplify Rank-1, Rank-5 as R-1 and R-5.

11. Implementation Details

In the student branch, we use ResNet50 [5] pre-trained on the ImageNet-1K dataset [3] as the default encoder backbone unless otherwise stated. The projection head consists of two fully connected layers with ReLU activation and a third linear layer with batch normalization [6]. For the contrastive training, we follow MoCo v2's [2] training style,

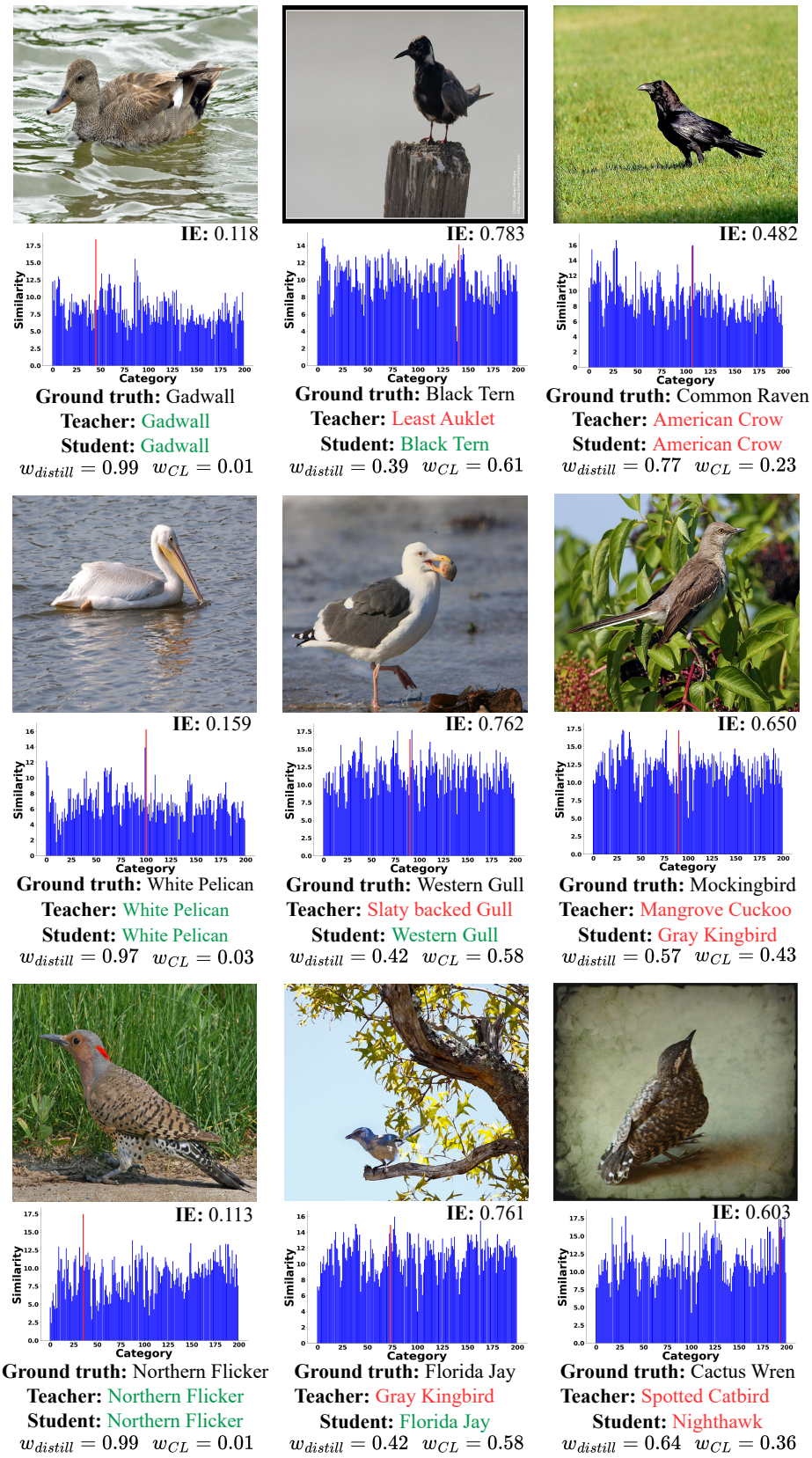


Figure 5. More examples of the prediction results and information entropy obtained by our method on the CUB dataset, and the probability distribution outputted by the teacher model are also illustrated. Predictions in red denote incorrect classifications, while predictions in green signify correct classifications.

Table 6. Descriptions of representative fine-grained categories generated by LLM using hand-crafted textual prompts. We limit the length of descriptions to 70 due to the input length limitation of CLIP's textual encoder. The descriptions provide rich semantic context for fine-grained representation learning.

Prompt: Describe the appearance of {fine-grained category} {coarse category}, with a word limit of no more than 70 words.		
Coarse category	**Fine-Grained category**	**Descriptions Generated by GPT-3**
bird	Pelagic Cormorant	The Pelagic Cormorant is a sleek seabird with a glossy black body and distinctive greenish-black iridescence. It features a long, slender neck and a sharp, pointed bill. During the breeding season, it displays striking white wing patches and a white throat patch. The eyes are bright blue, enhancing its unique look. Often seen near coastal waters, this cormorant is adept at diving for fish and is frequently spotted perched on rocky outcrops.
bird	Least Flycatcher	The Least Flycatcher is a small, slender bird with a brownish-olive back and pale underparts. It features a distinctively short, stubby bill and a faint eye ring. The wings are dark with two white wing bars, and its tail is relatively short. Its overall coloration is subtle, making it blend well into its surroundings. Often found in deciduous forests, this flycatcher is known for its quick, sharp calls and agile foraging behavior.
bird	White Pelican	The White Pelican is a large, graceful bird known for its striking all-white plumage and long wings. It has a massive, elongated bill with a distinctive pouch, ideal for catching fish. The head and neck are also white, with a slight yellowish tinge during breeding season. Its legs are short and webbed, and it has a rounded body. Often seen gliding over lakes and wetlands, the White Pelican is a majestic sight in flight.
car	Chevrolet TrailBlazer SS 2009	The 2009 Chevrolet TrailBlazer SS features a bold, muscular design that emphasizes its sporty character. Its aggressive front end showcases a prominent grille and sleek headlights, giving it a commanding presence. The sculpted body includes pronounced wheel arches and smooth lines, enhancing its athletic profile. At the rear, distinctive taillights and dual exhaust outlets highlight its performance pedigree. Overall, the TrailBlazer SS combines the utility of an SUV with a powerful, dynamic aesthetic.
car	Audi R8 Coupe 2012	The 2012 Audi R8 Coupe boasts an iconic, aerodynamic design with a low, wide stance. Its front features a bold Singleframe grille and sleek, angular headlights that convey aggression. The sculpted body showcases sharp lines and pronounced wheel arches, enhancing its sporty look. At the rear, distinctive taillights and a prominent diffuser accentuate its performance pedigree. Overall, the R8 combines luxury and cutting-edge design, embodying the essence of a high-performance supercar.
car	Lamborghini Aventador Coupe 2012	The 2012 Lamborghini Aventador Coupe features a striking, aerodynamic design with sharp angles and aggressive lines. Its low, wide stance is accentuated by prominent wheel arches and a bold front fascia with an angular grille. The signature Y-shaped headlights and sleek rear taillights enhance its modern aesthetic. With butterfly doors that open upward and a sculpted body, the Aventador embodies a blend of luxury and high-performance, making it an iconic supercar.
aircraft	Beechcraft 1900	The Beechcraft 1900 aircraft features a compact, twin-engine design with a rounded nose and a high-mounted, slightly swept-back wing. Its fuselage is sleek and narrow, accommodating a spacious cabin with large windows for passenger comfort. The two turboprop engines are mounted on the wings, enhancing its distinctive profile. Often painted in various airline liveries, the Beechcraft 1900 combines functionality and modern aesthetics, making it popular for regional and commuter flights.
aircraft	BAE 146-300	The BAE 146-300 aircraft features a sleek, wide-body fuselage with a rounded nose and low-mounted wings. Its distinctive four turbofan engines are mounted high on the wings, enhancing its unique profile. The wings are slightly swept back and equipped with winglets for improved aerodynamics. The spacious cabin includes large windows for passenger comfort. Often adorned in vibrant airline liveries, the BAE 146-300 combines modern design with functionality, making it ideal for regional and short-haul flights.
aircraft	SR-20	The SR-20 features a sleek, low-wing design with a rounded nose and a spacious, modern cockpit. Its elongated fuselage is complemented by large, oval windows that provide excellent visibility. The aircraft is powered by a single engine mounted at the front, giving it a streamlined profile. With a T-tail configuration and often finished in vibrant colors, the SR-20 combines advanced aerodynamics with a stylish appearance, making it a popular choice for general aviation.

and the momentum value and memory size are set to 0.999 and 65536 respectively. We set the batch size to 64 and use the SGD optimizer with a learning rate of 0.03, a momentum of 0.9, and a weight decay of 0.0001. During the training phase, the images were resized to 224×224 pixels, and the training epoch is set to 100. During the testing phase, images are first resized to 256×256 pixels and then center-cropped to 224×224 pixels.

References

[1] Qi Bi, Wei Ji, Jingjun Yi, Haolan Zhan, and Gui-Song Xia. Cross-level multi-instance distillation for self-supervised fine-grained visual categorization. *arXiv preprint arXiv:2401.08860*, 2024. 4

[2] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 4

[3] 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*, pages 248–255. Ieee, 2009. 4

[4] Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*, 2021. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[6] Sergey Ioffe. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 4

[7] Zhe Ma, Jianfeng Dong, Shouling Ji, Zhenguang Liu, Xuhong Zhang, Zonghui Wang, Sifeng He, Feng Qian, Xiaobo Zhang, and Lei Yang. Let all be whitened: Multi-teacher distillation for efficient visual retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4126–4135, 2024. 1

[8] Yangyang Shu, Anton Van den Hengel, and Lingqiao Liu. Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11392–11401, 2023. 4

[9] Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1365–1374, 2019. 1

[10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 3

[11] Zihu Wang, Lingqiao Liu, Scott Ricardo Figueroa Weston, Samuel Tian, and Peng Li. On learning discriminative features from synthesized data for self-supervised fine-grained visual recognition. In *European Conference on Computer Vision*, pages 101–117. Springer, 2025. 4