

Neural network-based discriminative feature formation and aggregation for fine-grained image classification*

Artem Kusakov^{1,*}, Andrii Oliinyk¹, Tetiana Kolpakova¹ and Oleksii Lukovka¹

¹ National University "Zaporizhzhia Polytechnic", 64 Universytetska St., Zaporizhzhia, 69011, Ukraine

Abstract

This paper presents a lightweight method for multi-level feature aggregation in fine-grained image classification. The proposed method is implemented within a compact convolutional neural network that learns hierarchical feature representations from input images and improves class separability by combining features extracted at multiple backbone depths. Specifically, feature maps from intermediate and deep convolutional blocks are aggregated via global average pooling, concatenated, and projected into a unified embedding space prior to classification. The model is fine-tuned end-to-end using gradient-based optimization with a standard cross-entropy objective. Experiments on two publicly available fine-grained datasets (CUB-200-2011 and Oxford Flowers 102) show that the proposed strategy consistently improves classification accuracy over a baseline CNN without multi-level feature aggregation. The obtained results indicate that multi-level feature aggregation within a compact CNN framework can improve fine-grained classification performance under controlled experimental conditions while preserving architectural simplicity and introducing only limited additional model complexity.

Keywords

fine-grained classification, convolutional neural network, multi-level feature aggregation, hierarchical feature representation, discriminative embedding, global average pooling, image recognition, deep learning 1

1. Introduction

Deep learning has become a dominant paradigm in computer vision, enabling substantial progress in image recognition and classification tasks. Convolutional neural networks (CNNs) learn hierarchical feature representations directly from raw image data and have substantially improved performance on large-scale visual recognition benchmarks.

Despite these advances, fine-grained image classification remains challenging. In such scenarios, object categories exhibit high inter-class similarity and subtle visual differences, making discriminative representation learning a critical factor in performance. Moreover, deep neural networks often rely on large annotated datasets and high model capacity, which may lead to overfitting when training data are limited or highly homogeneous.

Unlike conventional image classification, which distinguishes semantically different object categories (e.g., birds, cars, or flowers), fine-grained image classification focuses on discriminating visually similar subcategories within the same general category, such as bird species or flower species. These tasks are characterized by high inter-class visual similarity together with potentially large intra-class variation caused by pose, viewpoint, illumination, or background. Consequently,

*ADT-ISMDDC 2026: Advanced Digital Technologies: Intelligent Systems, Modeling, and Data-Driven Control, September 22–24, 2026, Odesa, Ukraine

^{1*} Corresponding author.

✉ kusakov@hey.com (A. Kusakov); olejnikaa@gmail.com (A. Oliinyk); t.o.kolpakova@gmail.com (T. Kolpakova); lukerlexa@gmail.com (O. Lukovka)

ORCID 0009-0008-3590-8664 (A. Kusakov); 0000-0002-6740-6078 (A. Oliinyk); 0000-0001-8307-8134 (T. Kolpakova); 0009-0007-2012-8422 (O. Lukovka)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

successful recognition relies on identifying subtle discriminative visual cues in addition to higher-level semantic information, rather than relying solely on coarse object-level characteristics. Benchmark datasets such as CUB-200-2011 and Oxford Flowers 102 are widely regarded as representative fine-grained classification benchmarks because they require discrimination between closely related subcategories exhibiting subtle visual differences.

Recent research has emphasized that classification performance depends not only on network depth but also on the quality of learned feature representations. Multi-scale analysis and feature fusion strategies have been explored to capture complementary information and improve separability of visually similar categories. In particular, aggregating representations extracted at different stages of a CNN can enhance discriminative capability without introducing substantial additional model complexity. Such strategies aim to enrich the final representation without substantially increasing computational cost.

In this work, a compact CNN backbone is intentionally adopted in order to focus the analysis on the contribution of multi-level feature aggregation rather than on the capacity of substantially deeper architectures. Using a relatively simple backbone provides a controlled setting in which the effect of aggregating intermediate and deep convolutional representations can be evaluated more directly. Therefore, the proposed method is positioned as a lightweight multi-level feature aggregation approach for improving discriminative representations within a compact CNN, rather than as a replacement for modern high-capacity architectures. Although numerous studies have demonstrated the effectiveness of sophisticated recognition architectures, it remains difficult to determine which performance gains originate from increased model capacity and which arise from improvements in representation learning itself. For this reason, feature aggregation provides an attractive direction because it can potentially improve representation quality without fundamentally changing the underlying backbone structure.

Motivated by these observations, this paper investigates a lightweight multi-level feature aggregation strategy for fine-grained image classification within a compact CNN framework. Rather than introducing a fundamentally new recognition architecture, the study examines whether aggregating intermediate and deep convolutional representations can improve discriminative representation learning while preserving architectural simplicity. Experimental results demonstrate consistent improvements over the corresponding single-level baseline.

2. Related Work

Research in image recognition has been fundamentally transformed by the emergence of deep learning and convolutional neural networks (CNNs). Early breakthroughs demonstrated that deep CNN models trained on large-scale datasets substantially outperform traditional hand-crafted feature pipelines, establishing hierarchical representation learning as the central paradigm in visual recognition [1]. Subsequent architectural developments, including VGG, Inception, and residual networks, further improved accuracy through increased depth and more effective optimization strategies [2]-[4]. Comprehensive overviews of deep learning and CNN-based visual understanding summarize these advances and highlight the importance of learned hierarchical representations in modern computer vision systems [5]-[7].

Beyond increasing network depth, numerous architectural refinements have been proposed to enhance representation quality and training stability. Dense connectivity patterns, widened residual blocks, and network-in-network structures aim to improve gradient flow, reduce redundancy, and increase parameter efficiency within convolutional backbones [8]-[10]. In parallel, research has explored automatic architectural optimization and evolutionary design strategies to further improve classification performance [11]. Lightweight CNN variants have also been developed for application-specific scenarios requiring efficiency under practical constraints [12]. These studies collectively

emphasize that representation quality and architectural design play a decisive role in visual recognition performance.

Fine-grained visual classification represents a particularly challenging subdomain of image recognition, where object categories exhibit high inter-class similarity and subtle discriminative cues. Benchmark datasets such as CUB-200-2011 and Oxford Flowers 102 exposed the limitations of generic CNN descriptors when applied to visually similar categories [13], [14]. Early fine-grained methods relied on part-based and region-aware modeling to localize informative object components, often incorporating region proposal mechanisms inspired by detection frameworks [15], [16]. However, such approaches may require additional annotations or introduce increased computational complexity.

To reduce dependence on explicit part supervision, later methods focused on strengthening global feature representations. Bilinear CNN models demonstrated that second-order feature interactions can significantly enhance fine-grained discrimination [17], while other approaches achieved competitive performance without explicit part annotations by improving global descriptor learning [18]. Recent work has also explored multi-view active recognition as a means of acquiring complementary information for fine-grained classification [19]. Broader studies and surveys systematically summarize the development of fine-grained recognition methods and underline the importance of representation-level improvements in high-similarity classification scenarios [20], [21].

In addition to deep architectures, classical multiresolution and texture-based analysis has long contributed to visual recognition theory. Wavelet-based multiresolution decomposition provides a theoretical foundation for capturing hierarchical spatial patterns [22], while local binary patterns and statistical texture models demonstrated the effectiveness of localized feature representations [23], [24]. Scattering convolution networks further combined predefined multiscale filters with learning-based processing to achieve stability and robustness to deformations [25]. These foundational ideas reinforce the importance of multi-level and multi-scale representation strategies in discriminative visual modeling.

Overall, existing research indicates that fine-grained recognition depends strongly on the ability to retain information across multiple levels of abstraction. Feature fusion strategies address this requirement by integrating representations extracted at different stages of a network. Recent methods have further improved fine-grained recognition through attentive pairwise feature interaction [26], channel-attention-based progressive multi-granularity training [27], Transformer-based architectures with attention-guided token selection [28], and semantic and attentive data mixing augmentation [29]. Although these approaches achieve strong recognition performance, they generally employ substantially deeper pretrained backbones, attention mechanisms, multi-stage training procedures, or more complex architectural components. The present work takes a more constrained approach by examining whether intermediate and deep convolutional features can be aggregated within a compact single-backbone CNN using only global average pooling, concatenation, and a linear projection layer. In this way, the study focuses on the representational effect of multi-level feature aggregation rather than on introducing a complex recognition framework.

3. Method

This study employs a convolutional neural network (CNN) as the core framework for discriminative image representation learning and classification. CNN architectures enable hierarchical feature extraction directly from raw images through stacked convolutional, pooling, and nonlinear activation layers. Such hierarchical processing allows the model to capture both local structural

details and high-level semantic patterns, which is particularly important for fine-grained recognition tasks involving visually similar categories.

The backbone follows a compact standard CNN design with stacked convolutional blocks and max pooling to progressively extract hierarchical visual features. This relatively simple architecture provides a controlled setting for evaluating the contribution of the proposed aggregation strategy without relying on substantially deeper backbone designs. Convolution kernels of size 3×3 are used throughout the backbone to maintain parameter efficiency while preserving local spatial information. Each convolutional block consists of a convolutional layer followed by batch normalization and a ReLU activation function, with max pooling applied where indicated.

To improve feature separability in fine-grained recognition tasks, the proposed aggregation strategy explicitly combines features extracted at multiple depths of the CNN backbone. Different convolutional layers capture complementary visual characteristics: intermediate layers encode localized texture and structural cues, whereas deeper layers represent higher-level semantic patterns. Therefore, multi-level feature representations are extracted and integrated prior to the final classification stage.

The resulting architecture integrates multi-level features while preserving the original backbone design, as illustrated in Figure 1.

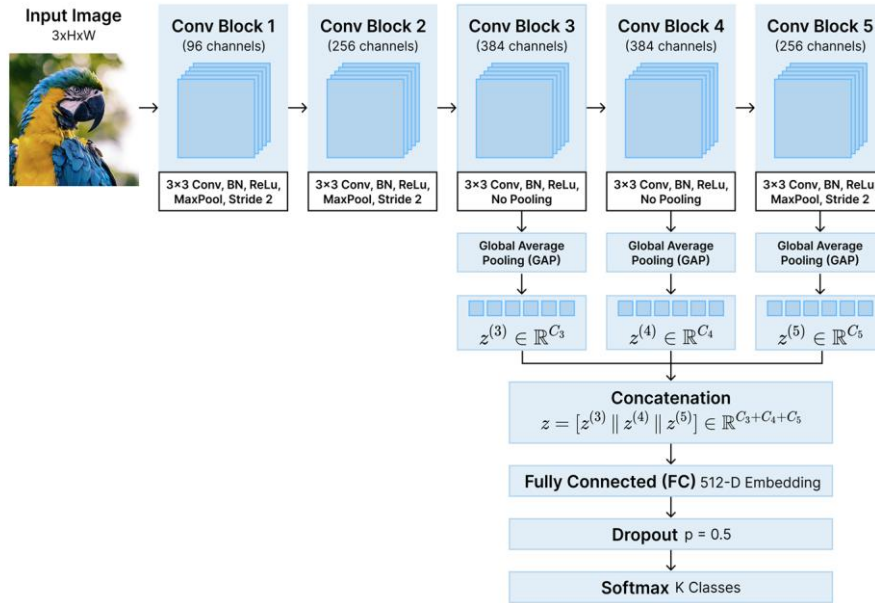


Figure 1: Architecture of the proposed multi-level feature aggregation framework.

The complete network architecture and layer configuration are summarized in Table 1.

Table 1
Architecture of the proposed convolutional neural network

Layer	Configuration
Conv Block 1	96 filters, 3×3 kernel, stride 2, batch normalization, max pooling
Conv Block 2	256 filters, 3×3 kernel, batch normalization, max pooling

Conv Block 3	384 filters, 3×3 kernel, batch normalization
Conv Block 4	384 filters, 3×3 kernel, batch normalization
Conv Block 5	256 filters, 3×3 kernel, batch normalization, max pooling
Multi-level Aggregation	GAP on Conv3–Conv5 feature maps + concatenation
Projection Layer	Fully connected, 512-dimensional embedding
Classification Layer	K neurons, softmax

The aggregation process can be formally described as follows. Let the output feature maps of the third, fourth, and fifth convolutional blocks be denoted as $F^{(3)}$, $F^{(4)}$ and $F^{(5)}$, where each $F^{(i)} \in \mathbb{R}^{C_i \times H_i \times W_i}$ represents a three-dimensional tensor containing C_i channels with spatial dimensions $H_i \times W_i$.

To obtain compact vector representations from each hierarchical level, global average pooling (GAP) is applied to the corresponding feature maps. The pooled feature vector of level i is computed as

$$z^{(i)} = \text{GAP}(F^{(i)}) = \frac{1}{H_i W_i} \sum_{u=1}^{H_i} \sum_{v=1}^{W_i} F^{(i)}(:, u, v), \quad (1)$$

where $z^{(i)} \in \mathbb{R}^{C_i}$ denotes the aggregated channel-wise descriptor obtained from the i -th convolutional block, for $i \in \{3, 4, 5\}$.

The multi-level discriminative representation is formed by concatenating the pooled vectors from the selected layers:

$$z = [z^{(3)} \parallel z^{(4)} \parallel z^{(5)}], \quad (2)$$

where $\parallel$ denotes vector concatenation and $z \in \mathbb{R}^{C_3+C_4+C_5}$. In the adopted architecture, this produces a 1024-dimensional concatenated descriptor before projection.

To ensure dimensional consistency, the concatenated representation is projected into a unified embedding space through a linear transformation:

$$\tilde{z} = W_f z + b_f, \quad (3)$$

where $W_f \in \mathbb{R}^{D \times (C_3+C_4+C_5)}$ and $b_f \in \mathbb{R}^D$ are learnable parameters of the projection layer, and $D = 512$ denotes the dimensionality of the unified embedding space. The resulting vector $\tilde{z} \in \mathbb{R}^D$ constitutes the fused discriminative embedding used for classification. A dropout layer with probability $p = 0.5$ is applied to the fused embedding prior to the final classification layer.

The projected representation is passed to the final classification layer followed by a softmax activation to produce class probabilities. The network is trained using the standard cross-entropy loss:

$$L = - \sum_{c=1}^K y_c \log \hat{y}_c, \quad (4)$$

where K is the number of target classes, y_c is the ground-truth label, and $\hat{y}_c$ denotes the predicted probability for class c .

4. Experiments

All experiments were implemented in Python using the PyTorch framework and executed on a workstation equipped with an NVIDIA GPU. The convolutional backbone was initialized with ImageNet-pretrained weights, while the newly introduced projection and classification layers were randomly initialized. The complete network was then fine-tuned end-to-end on each target dataset. Batch normalization layers were optimized jointly with the convolutional backbone during training. The same training protocol, preprocessing pipeline, optimizer settings, and input resolutions were used for both the baseline CNN and the proposed aggregation model unless stated otherwise.

The proposed method was evaluated on two publicly available fine-grained image classification benchmarks: CUB-200-2011 and Oxford Flowers 102. CUB-200-2011 contains 11,788 images belonging to 200 bird species characterized by high inter-class similarity and subtle discriminative cues, whereas Oxford Flowers 102 consists of 8,189 images from 102 flower categories with considerable appearance variability. The official dataset splits were used throughout all experiments. For CUB-200-2011, the official train/test split was used, while for Oxford Flowers 102, the standard training, validation, and testing partitions were preserved according to the official dataset protocol. The selected datasets provide complementary evaluation scenarios for fine-grained recognition. CUB-200-2011 emphasizes subtle inter-species variations among bird categories, whereas Oxford Flowers 102 contains substantial variation in shape, texture, and color across flower classes. Evaluating the proposed method on both datasets allows the proposed aggregation strategy to be assessed under different forms of visual similarity and intra-class variability.



Figure 2: Representative examples of visually similar categories from the CUB-200-2011 and Oxford Flowers 102 datasets.

Figure 2 illustrates the high inter-class similarity characteristic of both datasets. The selected bird categories have a similar overall appearance and differ mainly in subtle morphological characteristics, whereas the flower categories share comparable color and shape patterns while

differing in finer structural details. These examples illustrate why fine-grained recognition requires sensitivity to subtle class-specific visual differences.

During training, input images were resized and randomly cropped to the target input resolution, followed by random horizontal flipping and normalization using ImageNet mean and standard deviation statistics. During validation and testing, images were resized to the corresponding target resolution and normalized using the same statistics without random augmentation. For each target dataset, the network was fine-tuned for 200 epochs using stochastic gradient descent with momentum 0.9 and weight decay $1e-4$. The initial learning rate was set to 0.01 and reduced by a factor of 0.1 at epochs 100 and 150. A batch size of 32 was used. Dropout regularization with probability $p = 0.5$ was applied to the projected 512-dimensional embedding before the final classification layer to reduce overfitting. The selected training protocol was intentionally kept identical for both the baseline and the proposed model in order to ensure a fair comparison. By maintaining the same optimization strategy, learning rate schedule, regularization settings, and data preprocessing pipeline, the experimental evaluation isolates the influence of the proposed multi-level feature aggregation strategy from other factors that may affect classification performance. This controlled setup allows performance differences to be attributed primarily to the representation learning strategy rather than to variations in training conditions.

To assess training stability, the proposed model was fine-tuned three times on each dataset using different random seeds while keeping the remaining training protocol unchanged. The reported classification accuracies correspond to the mean and standard deviation across these independent runs.

To evaluate the effectiveness of the proposed multi-level feature aggregation strategy, a baseline CNN configuration without multi-level feature aggregation was implemented. In the baseline model, only the globally pooled output representation of the final convolutional block was projected into the same 512-dimensional embedding space and forwarded to the classification layer, whereas the proposed framework aggregates intermediate and deep convolutional representations as described in Section 3.

Table 2 summarizes the classification accuracy obtained by the baseline and the proposed model on both benchmark datasets.

Table 2

Classification accuracy (%) of the baseline CNN and the proposed multi-level aggregation model

Model	CUB-200-2011	Oxford Flowers 102
Baseline CNN	73.2	86.8
Proposed (multi-level aggregation)	76.1 ± 0.3	89.5 ± 0.2

As shown in Table 2, the proposed model improves the baseline by 2.9 percentage points on CUB-200-2011 and by 2.7 percentage points on Oxford Flowers 102. The relatively low standard deviations of 0.3 and 0.2 percentage points, respectively, indicate that the proposed aggregation strategy provides stable performance across independent training runs.

The purpose of this comparison is to isolate the contribution of multi-level feature aggregation under identical training conditions rather than to compete with substantially deeper architectures.

To quantify the contribution of individual aggregation stages, an ablation analysis was conducted by progressively integrating convolutional representations extracted from different depths of the network.

Table 3
Ablation study of multi-level feature aggregation

Model Variant	CUB (%)	Flowers (%)
Conv5 only (baseline)	73.2	86.8
Conv4 + Conv5	74.6	88.0
Conv3 + Conv4 + Conv5 (proposed)	76.1	89.5

The results demonstrate that incorporating intermediate convolutional representations consistently improves classification performance. The progressive gains suggest that intermediate representations provide complementary information beyond the final Conv5 features.

Additional experiments were performed to evaluate the influence of input resolution on classification performance. The proposed aggregation framework was tested using multiple input sizes while preserving identical training settings.

Table 4
Influence of input resolution on classification accuracy

Resolution	CUB (%)	Flowers (%)
224×224	76.1	89.5
280×280	77.3	90.4
336×336	78.1	91.0

Increasing the input resolution produces moderate but consistent improvements on both datasets. This behavior is expected in fine-grained classification tasks, where subtle local structures and texture patterns benefit from higher spatial resolution. The consistent gains across input sizes indicate that the aggregation strategy remains effective as additional spatial detail becomes available.

To place the obtained results in the context of existing fine-grained recognition research, Table 5 compares the proposed method with representative part-based, bilinear, attention-based, augmentation-based, and Transformer-based approaches.

Table 5
Comparison of the proposed method with representative fine-grained image classification approaches.

Method	Backbone	Pretraining	Extra supervision	Main idea	CUB-200-2011 (%)	Flowers 102 (%)
Part-Based R-CNN [15]	AlexNet + R-CNN	Image Net	Parts + boxes	Part detection	73.9	—

B-CNN [17]	D-Net + M-Net	Image Net	None	Bilinear pooling	84.1	—
Krause et al. [18]	VGG-19	Image Net	Bounding boxes	Automatic part generation	82.0	—
API-Net [26]	DenseNet-161	Image Net	None	Attentive pairwise feature interaction	90.0	—
CA-PMG [27]	ResNet-101	Image Net	None	Channel attention and progressive multi-granularity training	89.2	—
TransFG [28]	ViT-B/16	Image Net-21K	None	Transformer with attention-guided token selection	91.7	—
SADMix [29]	ResNet-50	Image Net	None	Semantic and attentive data mixing augmentation	89.2	—
Proposed	Compact CNN	Image Net	None	Multi-level GAP aggregation	76.1 ± 0.3	89.5 ± 0.2

Note: “None” indicates that no additional supervision beyond image-level class labels is used. A dash (“—”) indicates that the corresponding result was not reported in the original publication.

Table 5 shows that representative recent fine-grained recognition methods report higher absolute classification accuracy, although these results are obtained using different backbone architectures, pretraining datasets, input resolutions, and training protocols. In contrast, the proposed method focuses on a lightweight CNN architecture with simple multi-level feature aggregation based on global average pooling and feature concatenation, providing an architecturally simpler alternative while consistently improving performance over the corresponding baseline.

Consistent with the comparison presented in Table 5, the proposed approach introduces only additional global average pooling operations, feature concatenation, and an expanded linear projection while leaving the convolutional backbone unchanged.

Compared with the baseline configuration, the proposed aggregation strategy increases the projection input dimensionality from 256 to 1024, introducing 393,216 additional trainable weights in the projection layer. Since global average pooling and feature concatenation introduce no trainable parameters, the additional model complexity is confined to this single linear projection, while the convolutional backbone remains unchanged. Therefore, the proposed method achieves improved discriminative performance with only a modest increase in model complexity.

Despite the positive results, several limitations should be acknowledged. The proposed experiments were conducted using a compact CNN backbone initialized with ImageNet-pretrained weights and fine-tuned under controlled conditions. Consequently, the study does not evaluate the proposed method in combination with substantially deeper convolutional architectures, alternative pretraining schemes, or transformer-based feature extractors. In addition, only two fine-grained benchmark datasets were considered. Further investigation on larger and more diverse datasets would provide a more comprehensive assessment of the generalization capability of the proposed approach.

5. Conclusion

This paper investigated a lightweight multi-level feature aggregation method for fine-grained image classification within a compact convolutional neural network framework. The proposed approach integrates intermediate and deep convolutional representations through global average pooling and concatenation without relying on attention mechanisms, explicit part localization, external supervision, or complex multi-branch designs.

Experimental evaluation on the CUB-200-2011 and Oxford Flowers 102 datasets demonstrated that the proposed aggregation strategy consistently improved classification accuracy compared with the baseline CNN configuration. Additional ablation experiments confirmed the complementary contribution of intermediate convolutional representations, while resolution analysis showed stable performance improvements at higher input sizes.

Although deeper backbone architectures may achieve higher absolute recognition accuracy, the proposed method provides a compact alternative for improving feature separability without introducing substantial additional model complexity. The obtained results suggest that multi-level feature aggregation is a useful strategy for improving compact CNN-based models in fine-grained recognition tasks involving high inter-class similarity.

Future work may investigate integration with deeper backbone architectures, alternative feature fusion strategies, and evaluation on additional fine-grained visual recognition benchmarks.

Acknowledgements

The work was carried out within the framework of the state-funded research project “Artificial intelligence methods and tools for analyzing, collecting, and processing information in decision-making systems” at the National University “Zaporizhzhia Polytechnic” funded by the Ministry of Education and Science of Ukraine (state registration No. 0126U000515).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.5 to improve language quality and perform grammar and spelling checks. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems*, volume 25, 2012, pp. 1097–1105.
- [2] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, *arXiv:1409.1556*, 2015. doi:10.48550/arXiv.1409.1556.
- [3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.
- [4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.

- [5] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (2015) 436–444. doi:10.1038/nature14539.
- [6] Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M. S. Lew, Deep learning for visual understanding: A review, *Neurocomputing* 187 (2016) 27–48. doi:10.1016/j.neucom.2015.09.116.
- [7] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects, *IEEE Transactions on Neural Networks and Learning Systems* 33 (12) (2022) 6999–7019. doi:10.1109/TNNLS.2021.3084827.
- [8] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.
- [9] S. Zagoruyko, N. Komodakis, Wide residual networks, in: *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, 2016, pp. 87.1–87.12. doi:10.5244/C.30.87.
- [10] M. Lin, Q. Chen, S. Yan, Network in network, *arXiv:1312.4400*, 2014. doi:10.48550/arXiv.1312.4400.
- [11] Y. Sun, B. Xue, M. Zhang, G. G. Yen, Evolving deep convolutional neural networks for image classification, *IEEE Transactions on Evolutionary Computation* 24 (2) (2020) 394–407. doi:10.1109/TEVC.2019.2916183.
- [12] M. Wiecek, J. Silka, M. Woźniak, S. Garg, M. M. Hassan, Lightweight convolutional neural network model for human face detection in risk situations, *IEEE Transactions on Industrial Informatics* 18 (7) (2022) 4820–4829. doi:10.1109/TII.2021.3129629.
- [13] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The Caltech-UCSD Birds-200-2011 dataset, Technical Report, California Institute of Technology, Pasadena, CA, USA, 2011.
- [14] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, Bhubaneswar, India, 2008. doi:10.1109/ICVGIP.2008.47.
- [15] N. Zhang, J. Donahue, R. Girshick, T. Darrell, Part-based R-CNNs for fine-grained category detection, in: *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, Springer, Cham, 2014, pp. 834–849. doi:10.1007/978-3-319-10590-1_54.
- [16] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (6) (2017) 1137–1149. doi:10.1109/TPAMI.2016.2577031.
- [17] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1449–1457. doi:10.1109/ICCV.2015.170.
- [18] J. Krause, H. Jin, J. Yang, L. Fei-Fei, Fine-grained recognition without part annotations, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015, pp. 5546–5555. doi:10.1109/CVPR.2015.7299194.
- [19] R. Du, W. Yu, H. Wang, T.-E. Lin, D. Chang, Z. Ma, Multi-view active fine-grained visual recognition, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 1568–1578. doi:10.1109/ICCV51070.2023.00151.

- [20] Y. Kang, G. Chao, X. Hu, Z. Tu, D. Chu, Deep learning for fine-grained image recognition: A comprehensive study, in: Proceedings of the 4th Asia Pacific Information Technology Conference (APIT '22), 2022, pp. 31–39. doi:10.1145/3512353.3512359.
- [21] X.-S. Wei, Y.-Z. Song, O. Mac Aodha, J. Wu, Y. Peng, J. Tang, J. Yang, S. Belongie, Fine-grained image analysis with deep learning: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (11) (2022) 8927–8948. doi:10.1109/TPAMI.2021.3126648.
- [22] S. G. Mallat, A theory for multiresolution signal decomposition: The wavelet representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 11 (7) (1989) 674–693. doi:10.1109/34.192463.
- [23] T. Ojala, M. Pietikäinen, T. Mäenpää, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (7) (2002) 971–987. doi:10.1109/TPAMI.2002.1017623.
- [24] M. Varma, A. Zisserman, A statistical approach to texture classification from single images, International Journal of Computer Vision 62 (1–2) (2005) 61–81. doi:10.1023/B:VISI.0000046589.39864.ee.
- [25] J. Bruna, S. Mallat, Invariant scattering convolution networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1872–1886. doi:10.1109/TPAMI.2012.230.
- [26] P. Zhuang, Y. Wang, Y. Qiao, Learning attentive pairwise interaction for fine-grained classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 13130–13137. doi:10.1609/aaai.v34i07.7016.
- [27] P. Zhao, Q. Miao, H. Yao, X. Liu, R. Liu, M. Gong, CA-PMG: Channel attention and progressive multi-granularity training network for fine-grained visual classification, IET Image Processing 15 (14) (2021) 3718–3727. doi:10.1049/ipr2.12238.
- [28] J. He, J.-N. Chen, S. Liu, A. Kortylewski, C. Yang, Y. Bai, C. Wang, TransFG: A transformer architecture for fine-grained recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 852–860. doi:10.1609/aaai.v36i1.19967.
- [29] M. He, Q. Cheng, G. Qi, Weakly supervised semantic and attentive data mixing augmentation for fine-grained visual categorization, IEEE Access 10 (2022) 35814–35823. doi:10.1109/ACCESS.2022.3163302.