Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ28608

Beyond Accuracy: A CNN Baseline and Failure Analysis for Landmark Classification on WikiScenes

Julia Gao

Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, USA
Juliagao10405@gmail.com

Abstract. Landmark classification remains a challenging task in computer vision due to high intra-class variation, long-tailed distributions, and real-world conditions such as occlusion and changes in viewpoint. This work introduces a systematic RGB-only formulation of the WikiScenes dataset, reframed as a 99-class benchmark that enables controlled evaluation of landmark recognition without relying on multimodal cues. We evaluate ResNet-18 and ResNet-50 across six distinct training regimes, including fine-tuning, training from scratch, and using frozen backbones. To address class imbalance, we apply weighted sampling and data augmentation, and assess performance using top-1 accuracy, micro-mAP*, and macro-mAP. Fine-tuned pretrained models consistently outperform other approaches. ResNet-18 achieves the highest micro-mAP of 0.8364 and macro-mAP of 0.7120, while ResNet-50 reaches the best top-1 accuracy at 84.3 percent. PCA embeddings and confusion matrices reveal that fine-tuned models produce compact and separable representations, whereas weaker training regimes result in diffuse and overlapping clusters. By establishing WikiScenes as a reproducible RGB-only benchmark and analyzing systematic failure modes, this study provides new insights into the limitations of image-only convolutional neural networks.

Keywords: landmark classification, convolutional neural networks, transfer learning, RGB image benchmark, computer vision

1. Introduction

Understanding visual content is central to applications such as photo sharing, digital assistants, mapping services, and augmented reality. Linking images to geographic context enables photo organization, recommendation, and spatial retrieval [1, 2], but GPS metadata is often unavailable due to privacy constraints or device limitations [3]. Landmark recognition, which involves classifying geographic sites in images, plays a critical role in tagging, travel recommendation, and city-scale reconstruction [4-6]. Despite progress driven by large-scale benchmarks like Google Landmarks v2 (GLDv2) [7], the task remains challenging due to extreme variation in viewpoint, scale, lighting, and occlusion—factors that continue to hinder the performance of pretrained convolutional neural networks [8-10]. These persistent challenges highlight the need for controlled,

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.TJ28608

reproducible benchmarks that isolate visual recognition performance and expose the limitations of current CNN-based approaches.

To address this, we introduce a novel benchmark: an RGB-only classification benchmark on the WikiScenes dataset [11], reframed as a 99-class challenge to systematically evaluate pretrained CNNs without multimodal cues. We conduct a comprehensive study of ResNet-18 and ResNet-50 across six training regimes, including fine-tuning, scratch, and frozen-backbone configurations. In addition to quantitative metrics, we analyze PCA embeddings and confusion matrices reveal where CNNs succeed and where they fail. This benchmark delivers a reproducible baseline and new insights into the limitations of RGB-only models, motivating future multimodal and adaptive approaches to unconstrained landmark recognition.

2. Related work

2.1. Deep learning for image classification

Deep convolutional networks have driven progress in image classification. The ResNet family introduced residual connections that enabled training of very deep models and remain widely used as transfer learning backbones [12]. DenseNet improved gradient flow with dense connectivity [13]. Later architectures such as EfficientNet [14], ConvNeXt [15], and Vision Transformers (ViT) [16] advanced accuracy and efficiency through scaling and attention-based design, with lightweight variants such as DeiT reducing data demands [17]. In parallel, self-supervised approaches such as SimCLR [18] and BYOL [19] broadened pretraining strategies beyond supervised ImageNet [20], which continues to serve as the standard foundation for downstream vision tasks. Visualization and interpretability studies such as Zeiler and Fergus [21] provided insight into CNN behavior, while augmentation techniques including RandAugment [22], Hide-and-Seek [23], Cutout [24], and Mixup [25] improved robustness under weak supervision.

2.2. Landmark recognition

Landmark recognition has been studied both as retrieval and classification. Early benchmarks such as Oxford5k and Paris6k evaluated handcrafted descriptors [4,5]. Deep architectures including NetVLAD [1], DELF [2], R2D2 [26], and DISK [27] improved retrieval through learned global and local features. Metric learning approaches such as ArcFace [24] and FaceNet [28] enhanced discriminability for large-scale recognition. The release of GLDv2 scaled the problem to hundreds of thousands of categories, driving progress in global–local feature aggregation [7]. Beyond retrieval, early classification-based methods such as Zhu et al. [29] explored hierarchical multimodal feature learning for landmark categorization. More recent multimodal datasets such as WikiScenes extended the task to include text and 3D geometry [11], while geometry-aware methods such as Halo-NeRF [30] introduced neural scene representations for structure-guided recognition. Despite these advances, pretrained CNN backbones degrade under real-world variation [8-10], motivating baseline studies of RGB-only classification.

2.3. Challenges in in-the-wild classification

Real-world landmark images show large intra-class variation due to viewpoint, occlusion, and clutter. Studies of occlusion sensitivity [21, 24] and augmentation [22, 23, 25] highlight the fragility of CNNs under these conditions. Benchmarks of distribution shift further demonstrate the gap between curated training data and natural imagery [8-10, 14]. Domain adaptation approaches such as

DOI: 10.54254/2755-2721/2026.TJ28608

DANN [8] aim to close this gap, but remain limited when applied at landmark scale. These findings underscore the need for systematic evaluation of pretrained CNNs under realistic landmark variation.

3. Method

We fine-tune ResNet-18 and ResNet-50 for 99-class landmark classification on WikiScenes [11]. The dataset, originally built for multimodal learning, provides diverse interior, exterior, and viewpoint images per landmark; here, all are pooled into single classes. Images are resized to 224×224 and augmented with flips, rotations, crops, and perspective distortion. The dataset is split 80/20 for training and validation using stratified sampling. To address the dataset's long-tailed distribution, we use PyTorch's WeightedRandomSampler to balance minibatches. Both models are fine-tuned end-to-end with the Adam optimizer (batch size 32), replacing the final fully connected layer with a 99-way classifier. Training proceeds for up to 150 epochs with early stopping and predictions are taken from the top softmax output.

4. Evaluation strategy

We evaluate models using both quantitative and qualitative analyses. Accuracy, micro-mAP*, and macro-mAP measure overall effectiveness and fairness under class imbalance. To probe representation quality, we project latent features with PCA and compute normalized confusion matrices to reveal per-class error patterns. Together, these methods provide a clear view of model performance, representation stability, and systematic failure modes.

5. Results

Table 1. Quantitative evaluation. Performance is reported as micro-mAP* (averaged across all samples), macro-mAP (average precision per class and then across classes), and top-1 accuracy

Model Variant	mAP*	mAP	Accuracy
ResNet-18 (w/ weights)	0.8364	0.7120	83.58
ResNet-18 (w/o weights)	0.5703	0.4601	69.32
ResNet-18 (final layer)	0.2295	0.2610	28.62
ResNet-50 (w/ weights)	0.8027	0.6974	84.28
ResNet-50 (w/o weights)	0.2637	0.2820	42.95
ResNet-50 (final layer)	0.1458	0.2329	18.78

ResNet-18 and ResNet-50 were evaluated under three regimes: fine-tuned, trained from scratch, and final-layer only (Table 1). Fine-tuned models consistently outperform the others. ResNet-18 achieves the highest micro-mAP (0.8364) and macro-mAP (0.7120), while ResNet-50 attains the best top-1 accuracy (84.3%). This contrast highlights a trade-off between balance and specialization: ResNet-50 excels at recognizing frequent classes, raising overall accuracy, but underperforms on rare categories, reducing mAP. ResNet-18 distributes performance more evenly across the long-tailed dataset, which not only improves fairness but also produces more stable convergence across training runs. Models trained from scratch fare poorly, especially ResNet-50, underscoring the difficulty of optimizing deeper networks without large-scale pretraining. Final-layer-only training is weakest, confirming that ImageNet features remain too generic for fine-grained landmark

DOI: 10.54254/2755-2721/2026.TJ28608

classification without deeper adaptation. Notably, the stronger performance of ResNet-18 on both mAP metrics suggests that smaller models can sometimes better accommodate class imbalance by avoiding overfitting to dominant categories. These findings indicate that successful landmark classification depends less on sheer depth than on effective transfer learning and training strategies tailored for imbalanced data.

6. Visual analysis

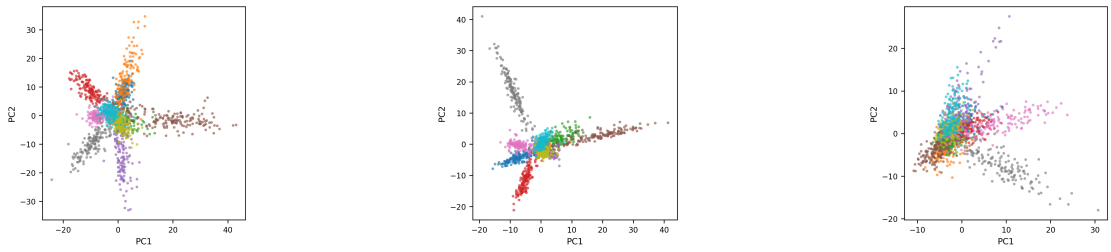

Figure 1. PCA embeddings for top 10 most frequent classes for ResNet-18 w/weights, ResNet-50 w/weights, and ResNet-18 w/o weights respectively

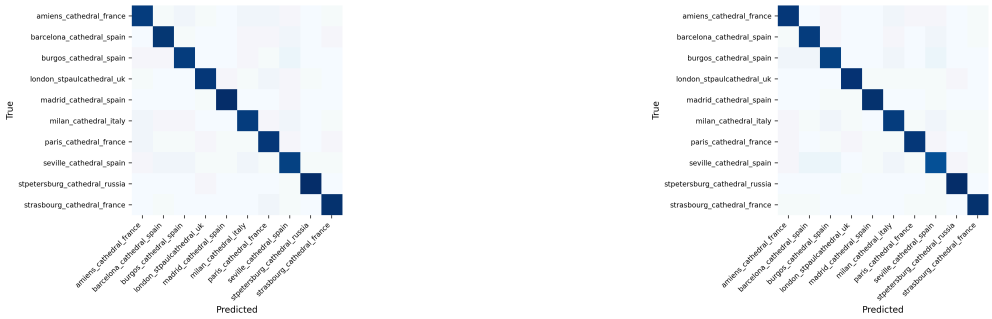

Figure 2. Normalized confusion matrix of top-10 most frequent classes using ResNet-18 w/weights and ResNet-50 w/weights respectively

The qualitative tools in Figures 1 and 2 provide insight into why these performance patterns emerge. PCA embeddings show that fine-tuned ResNet-18 and ResNet-50 form tight, well-separated clusters, demonstrating that pretrained features adapt effectively to landmarks and preserve class distinctions even across changes in viewpoint, scale, and lighting. These compact groupings indicate stable latent representations that generalize beyond the training set. In contrast, the scratch-trained ResNet-18 produces overlapping and diffuse clusters with no clear separation, suggesting unstable decision boundaries and explaining its low scores. ResNet-50 embeddings, while compact, show slightly less separation than ResNet-18, implying that deeper models may capture broader context but blur distinctions between visually similar landmarks. Confusion matrices provide complementary evidence about classification behavior. ResNet-18 shows a dark diagonal with minimal confusion while ResNet-50 exhibits a lighter diagonal with faint off-diagonal entries, indicating error among similar landmarks. This contrast reinforces the quantitative metrics: ResNet-18 achieves higher mAP and mAP* by distributing performance more evenly, while ResNet-50 gains in top-1 accuracy by prioritizing dominant classes. Together, these analyses show that fine-

DOI: 10.54254/2755-2721/2026.TJ28608

tuning not only improves accuracy but also reshapes latent spaces and error patterns, revealing interpretable failure modes.

7. Conclusion

We present the first RGB-only baseline for landmark classification on the WikiScenes dataset, reframed as a 99-class benchmark to isolate and evaluate visual recognition performance without auxiliary signals. This novel formulation enables systematic comparison of CNN training regimes, revealing that fine-tuned pretrained models consistently outperform scratch and frozen-backbone approaches. ResNet-50 achieves the highest top-1 accuracy, while ResNet-18 delivers stronger micro- and macro-mAP scores, highlighting a trade-off between depth and class balance in long-tailed distributions.

PCA embeddings and confusion matrices reinforce these findings, establishing WikiScenes as a reproducible benchmark for landmark classification that exposes the limitations of image-only CNNs. Future work may explore multimodal cues (e.g., 3D structure, text) and adaptive strategies (e.g., continual or few-shot learning) to advance robust, scalable landmark recognition in real-world conditions.

Acknowledgements

I would like to thank Rana Hanocka for her mentorship and guidance throughout the development of this work. Her insights and feedback were instrumental in shaping the direction and clarity of this research.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, ´ and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[2] Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[3] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2018.

[4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[5] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

[6] Filip Radenovic, Giorgos Tolias, and Ondrej Chum. Fine-tuning ´ cnn image retrieval with no human annotation. In IEEE Trans6 actions on Pattern Analysis and Machine Intelligence (TPAMI), 2018.

[7] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2: A large-scale benchmark for instance-level recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations (ICLR), 2019.

[9] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, Tony Lee, Etan Dave, Ian Stavness, Yixuan Guo, Bruce Earnshaw, Casey LewWilliams, Christopher Re, Percy S. Liang, Sarah Carmichael, David Gomez, Chris Piech,

DOI: 10.54254/2755-2721/2026.TJ28608

Dinesh Ramasamy, Jascha SohlDickstein, and Percy Liang. Wilds: A benchmark of in-thewild distribution shifts. In International Conference on Machine Learning (ICML), 2021

[10] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In International Conference on Machine Learning (ICML), 2019.

[11] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

[12] 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 (CVPR), pages 770–778, June 2016.

[13] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[14] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In International Conference on Machine Learning (ICML), 2019.

[15] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations (ICLR), 2021

[17] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve J ´ egou. Training data-efficient ´ image transformers & distillation through attention. In International Conference on Machine Learning (ICML), 2021.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International Conference on Machine Learning (ICML), 2020.

[19] Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin Tal- ´ lec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal ´ Valko. Bootstrap your own latent: A new approach to selfsupervised learning. In NeurIPS, 2020.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV), 115(3): 211–252, 2015.

[21] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In European Conference on Computer Vision (ECCV), 2014.

[22] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020.

[23] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: A data augmentation strategy for weakly-supervised localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[24] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv: 1708.04552, 2017.

[25] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In Proceedings of the International Conference on Learning Representations (ICLR), 2018.

[26] Jerome Revaud, Philippe Weinzaepfel, Martin Humenberger, and Yohann Cabon. R2d2: Reliable and repeatable detector and descriptor. In NeurIPS, 2019.

[27] Michal Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. In NeurIPS, 2020.

[28] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR, 2015.

[29] Lei Zhu, Jialie Shen, Hai Jin, Liang Xie, and Ran Zheng. Landmark classification with hierarchical multi-modal exemplar feature. IEEE Transactions on Multimedia, 17(7): 981–993, 2015.

[30] Chen Dudai, Morris Alper, Hana Bezalel, Rana Hanocka, Itai Lang, and Hadar Averbuch-Elor. Halo-nerf: Learning geometryguided semantics for exploring unconstrained photo collections. In Computer Graphics Forum (Eurographics), volume 43, page e15006, 2024.