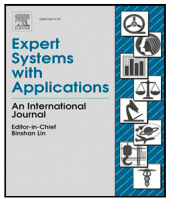
Expert Systems With Applications 234 (2023) 121018

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Efficient deep-narrow residual networks using dilated pooling for scene recognition☆

Zhinan Qiao [a], Xiaohui Yuan [a,*], Runmei Zhang [b], Tian Chen [c], Chaoning Zhang [d]

[a] University of North Texas, 3940 N. Elm, Denton, TX, 76207, USA
[b] Anhui Jianzhu University, Hefei, 230022, China
[c] Hefei University of Technology, Hefei, 230009, China
[d] Kyung Hee University, Seoul, 02447, South Korea

ARTICLE INFO

Keywords:
Convolutional Neural Network
Scene recognition
Efficient learning

ABSTRACT

This paper aims to address the challenges posed by complex scenery image classification. Most of the existing deep learning networks are trained and evaluated using ImageNet. However, when these models are applied to scenery images, dramatic performance degradation is observed due to the change in data characteristics. To challenge the prevailing practices in network design, we investigate the impact of altering data on the performance of deep networks. Specifically, we introduce a novel data-oriented network design to emphasize the importance of considering the unique characteristics of the data. Our proposed approach is a Deep-Narrow Network, which incorporates a Dilated Pooling module built upon the ResNet architecture. Compared to ResNet, our approach achieves a significant reduction of floating-point operations by 51.5% and in the number of parameters by 54.5%. Remarkably, despite the reduction in computational complexity and model size, our design exhibits a 0.4% increase in overall accuracy. This approach offers an efficient and effective means of scaling the network according to the data characteristics while maintaining highly competitive performance.

1. Introduction

Since the introduction of AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), there have been numerous advancements in deep Convolutional Neural Networks (CNNs). Researchers have explored different strategies to enhance network performance, such as increasing network depth or width. Deepening the networks by adding more convolutional layers has shown promising results (He, Zhang, Ren, & Sun, 2016; Simonyan & Zisserman, 2015; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), while widening the networks (i.e., increasing the number of channels in a CNN layer) has also gained traction. Zagoruyko and Komodakis (2016) proposed wide deep residual networks that outperformed ResNet (He et al., 2016) models, achieving state-of-the-art performance on ImageNet (Jia et al., 2009) and CIFAR (Krizhevsky, Sutskever, & Hinton, 2014). Similarly, Xie, Girshick, Dollár, Tu, and He (2017) introduced ResNeXt, which widened the residual blocks and utilized group convolution, resulting in improved performance on ImageNet-5K and COCO object detection datasets (Lin et al., 2014) compared to ResNet. Building upon these advancements, Zhang et al. (2022) introduced ResNeSt, which preserved the wider network layout and multi-branch strategy while incorporating a modulated architecture to enhance feature learning. The proposed ResNeSt networks demonstrated further improvements in performance on the ImageNet dataset.

Integrating multiple networks has been developed for scene recognition in the era of deep learning, enjoying wide adoption in the research community (Cheng, Lu, Feng, Yuan, & Zhou, 2018; Qiao, Yuan, & Elhoseny, 2020; Zhu, Deng, & Newsam, 2019). By concatenating or integrating network outputs, a unified scene feature representation is generated, resulting in improved performance (Cheng et al., 2018; Herranz, Jiang, & Li, 2016; Jia et al., 2009; Qiao et al., 2020; Selvaraju et al., 2017; Wang, Wang, Wang, Zhang, & Qiao, 2017; Xia, Zeng, Leng, & Fu, 2019; Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017a). In addition to network integration, various methods employing different algorithms and network architectures have been developed to tackle the scene recognition challenge (Gupta, Sharma, Dinesh, & Thenkanidiyoor, 2021; Lin et al., 2022; Lv, Dong, Zhang, & Xu, 2023; Rehman, Saleem, Khan, Jabeen, & Shafiq, 2021; Shi, Zhu, Yu, Wu, & Shi, 2019; Wang, Peng, & De Baets, 2020). These studies typically utilize pre-trained deep networks as feature extractors and aggregate them with other network structures to accomplish scene recognition tasks. Recent research has also introduced add-on modules to enhance performance,

☆ The source code is available at https://github.com/zn-qiao/deep-narrow-network. The corresponding authors include Xiaohui Yuan and Runmei Zhang.

* Corresponding author.
E-mail address: xiaohui.yuan@unt.edu (X. Yuan).

https://doi.org/10.1016/j.eswa.2023.121018
Received 24 April 2023; Received in revised form 9 July 2023; Accepted 17 July 2023
Available online 27 July 2023

such as adaptive learnable models that assign weights to different scales in scene recognition tasks (Qiao, Yuan, Zhuang, & Meyarian, 2021; Yuan, Qiao, & Meyarian, 2022). Despite these advancements, the design of the network backbone, which constitutes the core architecture, has received limited attention in the existing literature. This highlights the existing gap between the state-of-the-art generic deep neural network design and the design specifically tailored for scene recognition. Consequently, there is an urgent and imperative need to incorporate the most advanced network design theories and practices into the domain of scene recognition.

To address the challenges of classifying complex scenery images that contain multiple objects of varying sizes, we need to maximize both the depth and width of the network while considering computational constraints. However, using larger networks requires a substantial amount of training data, making it a less feasible option. Therefore, it is crucial to understand the roles of network layers and channels in order to determine whether deeper networks or more channels are more suitable for better comprehension of complex scenery images. Previous studies have explored this question.

Lu, Pu, Wang, Hu, and Wang (2017) argued that a combination of depth and width provides neural networks with expressive power. Tan and Le (2019) emphasized the importance of balancing network depth and width by maintaining a constant depth/width ratio, demonstrating its effectiveness on ResNet and MobileNet architectures. In addition to manually designed networks, Deep Neural Architecture Search techniques have been proposed to optimize network depth and width (Guo, Wang, Li, & Yan, 2020; Zoph & Le, 2017).

Most existing methods have been developed and evaluated using datasets like ImageNet (Jia et al., 2009) and CIFAR (Krizhevsky et al., 2014). These datasets primarily consist of object-centric images, where a single object dominates the image with a relatively homogeneous background. The focus is on recognizing the object itself. However, scenery images present a more complex view with multiple objects of different sizes and background clutter. Correctly classifying such images requires understanding the collective characteristics of many objects. This difference in dataset characteristics may lead to a bias in the design of convolutional neural networks (CNNs) towards object-centric features.

Deep networks with more layers are effective in extracting distinctive scale features through diverse receptive fields, while networks with more channels excel in capturing fine-grained patterns (Tan & Le, 2019). In many applications, both types of information are crucial for accurate image recognition, but their relative prominence is often overlooked.

This paper introduces a new neural network design strategy, focusing on the importance of learning the spatial layout of objects for a comprehensive understanding of a scene. Object-centric images typically contain a single object, where the spatial layout contributes less to the semantic meaning of the image. Due to subtle differences between objects, the detailed patterns and textures of individual objects become more representative. Networks that prioritize learning various features are better suited for object recognition tasks. Considering the distinct characteristics of scenery images, we hypothesize that, for scene recognition, learning the spatial information encoded in network layers has a greater impact on performance compared to learning channel-wise information.

In the following sections, we present comprehensive experimental results that demonstrate the benefits of deeper networks for scenery images, while the impact of altering network width is marginal. Building upon these observations, we propose a Deep-Narrow Network for complex scenery image classification. This network increases network depth while reducing its width, and incorporates a Dilated Pooling module to enhance the spatial and scale features of objects.

Our key contributions include the following:

1. We conducted an empirical analysis of the deep network design with respect to the prioritization of the data characteristics. For tasks such as scenery image recognition, learning the spatial layout is crucial, whereas tasks such as object recognition benefit more from learning various features.
2. From the perspective of image characteristics, we conducted an in-depth analysis of the impact of depth and width in CNNs on scene recognition tasks. Our extensive experiments revealed that deeper networks are more suitable for scene recognition compared to wider networks. Spatial information plays a vital role in scene recognition, enabling networks to better understand the overall spatial layout of a scene. This differs from object-centric images, where the detailed patterns and textures of objects are more significant.
3. To enhance spatial and scale features, we propose a Deep-Narrow Network, which incorporates a Dilated Pooling component. Through extensive experimentation, we demonstrate the significant benefits of deepening the network for scene images in terms of both efficiency and accuracy. Our approach provides valuable insights into scaling the network based on the specific characteristics of the data, allowing for improved performance without compromising efficiency.

The remainder of this paper is organized as follows. Section 2 reviews the network design strategies, with a specific focus on the network width and depth design. Additionally, we delve into the existing studies on scene recognition to provide a comprehensive understanding of the current state of the field. Section 3 presents the distinctive characteristics of scenery images and introduces our proposed Deep-Narrow Network with Dilated Pooling strategy. Section 4 discusses our experimental results including a comparison study using object classification and scene recognition datasets. Section 5 concludes our paper with a summary and directions for future research.

2. Related work

Deep & Wide Networks The depth of a network plays an important role in the success of convolutional neural networks (CNNs). As the network becomes deeper, it becomes capable of approximating the target function more precisely, resulting in improved performance. This significance of network depth was further highlighted by the success of VGGNet (Simonyan & Zisserman, 2015) and Inception (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017; Szegedy et al., 2015, 2016) in the ILSVRC competition. The introduction of ResNet (He et al., 2016), as a continuation of deeper networks, revolutionized the depth of deep networks by introducing residual learning and identity mapping concepts to CNNs. ResNet's innovative approach allows for extremely deep networks and has demonstrated enhanced performance in image recognition tasks. Recent studies have placed significant emphasis on advancing deep networks through various techniques. One notable approach involves the incorporation of dense shortcuts, as explored in the work of Zhang et al. (2021), which has shown improvements in gradient flow and overall performance for deep networks. Another technique gaining attention is the integration of Multi-Layer Perceptron (MLP) components into convolutional networks, as demonstrated by Li, Hassani, Walton, and Shi (2023) and Shen et al. (2023). These efforts exemplify the ongoing endeavors to refine and optimize deep network architectures for a wide range of applications. By integrating such techniques, there is potential to enhance the capabilities and effectiveness of deep learning models in tackling complex tasks.

Network width has also emerged as a crucial parameter in deep network design. Wide ResNet (Zagoruyko & Komodakis, 2016) introduced an additional factor to control the width of ResNet, and experimental results demonstrated that widening the network could be a more effective approach for improving performance compared to simply increasing its depth. Xception (Chollet, 2017), an extreme

version of the Inception architecture, employed wider structures by modifying the original inception block, leading to improved performance. ResNeXt (Xie et al., 2017) introduced the concept of cardinality to increase the width of ResNet, achieving remarkable results in the 2016 ILSVRC classification task. The success of ResNeXt has established the notion that widening deep networks is an effective strategy for enhancing model performance. ResNeSt (Zhang et al., 2022) further built upon the wide architecture of ResNeXt and achieved superior performance in image and object recognition tasks. Most recently, Chen et al. (2022) proposed WGNet, a graph convolutional neural network designed for 3D point cloud classification. WGNet incorporates local dilated connections and context-aware features to improve accuracy, emphasizing the importance of a wider network for capturing complex decision boundaries.

Effects of Depth and Width Although depth and width are proven to be essential parameters in network architecture design, the effect of depth and width, i.e., what deep and wide networks learn remains seldom explored. Most of the existing literature focuses on the effect of width and depth separately or the trade-off between depth and width in the network design (Lu et al., 2017). Nguyen and Hein (2018) suggest that having a hidden layer that is wide enough is crucial to ensure that the network is capable of creating separate decision boundaries. Veit, Wilber, and Belongie (2016) studied the mechanism behind ResNet and suggested deep residual networks have indicated that they might not be functioning as a single deep network, but instead, they could be operating as a collection of multiple relatively shallow networks. They implemented a group of similarly shallow networks and demonstrated that these networks perform better than the conventional deep residual networks. Tan and Le (2019) claimed that deep networks can make use of a larger receptive field while wide networks can better capture fine-grained features. Nguyen, Raghu, and Kornblith (2021) explored the effects of width and depth and found a characteristic structure named block structure. They demonstrated that for different models, the block structure is unique, but the representations outside the block structure trends to be similar despite the setting of depth and width. In our paper, we analyze the effect of depth and width in CNNs from the perspective of image characteristics.

In recent years, there has been notable progress in the theoretical analysis of deep and wide networks across different sub-domains. Mirzadeh et al. (2022) conducted an investigation into the relationship between network width and catastrophic forgetting. They proposed a hypothesis emphasizing the importance of high orthogonality among gradients from different tasks, which is induced by wider networks, increased gradient sparsity, and a lazy training regime. Bordelon and Pehlevan (2022) introduced a path integral formulation to study the gradient flow dynamics in infinite-width networks within the feature learning regime. They developed a polynomial-time numerical procedure to solve the saddle point equations for deep networks and demonstrated, through numerical experiments, that the solutions provide valuable insights into network training across different feature learning strengths, widths, and depths. Their theory was compared to various approximate methods, including perturbation theory.

In a similar vein, Radhakrishnan, Belkin, and Uhler (2023) explored the benefits of network depth and highlighted that very deep networks can achieve optimality with careful selection of activation functions. They established that deep networks with activations such as ReLU or tanh do not reach optimality. Furthermore, they discussed the advantages of using infinitely wide and deep networks for classification tasks compared to regression settings, where these networks are far from optimal. Despite the increasing focus on theoretical analysis of deep and wide networks, limited research has been conducted on investigating their effects based on data preference.

Scene Recognition The integration of multiple networks has been widely adopted in the scene recognition community as a pioneering approach in the era of deep learning (Cheng et al., 2018; Qiao et al., 2020; Zhu et al., 2019). These methods typically concatenate or integrate the outputs of multiple networks to generate a unified scene feature representation, leading to improved performance (Cheng et al., 2018; Herranz et al., 2016; Jia et al., 2009; Qiao et al., 2020; Selvaraju et al., 2017; Wang et al., 2017; Xia et al., 2019; Zhou et al., 2017a). While some studies have devised trainable integration methods (Seong et al., 2019; Seong, Hyun, & Kim, 2020), the backbone network architectures remain unchanged.

In addition to network integration, there are alternative methods that employ different algorithms and network architectures to tackle the scene recognition challenge (Gupta et al., 2021; Rehman et al., 2021; Shi et al., 2019; Wang et al., 2020). For instance, Wang et al. (2020) proposed an "adaptive discriminative metric learning" (DFF-ADML) method to address the complexity of scene images. DFF-ADML utilized pre-trained CNNs to extract multiple features from training examples and employed linear transformations to map these features into a common space, preserving more information about the scene images. This adaptive fusion of features enhanced the preservation of comprehensive content and improved scene recognition performance. Another study by Rehman et al. (2021) combined the conventional Bag of Visual Words (BOVW) technique with AlexNet by fusing the features extracted from both models. They extensively experimented with different BOVW sizes and carefully designed and tested their pipeline components, resulting in an effective scene recognition method. Similarly, Gupta et al. (2021) utilized pre-trained CNNs for feature extraction and proposed a threshold-based approach to select the most prominent features and filter out less informative ones. They also integrated a grouping-based method to remove redundant features, leading to improved performance. Furthermore, Lin et al. (2022) leveraged two pre-trained CNNs to extract original feature maps and employed techniques such as class activation mapping to identify salient regions and extract local scene features. They also incorporated bidirectional long short-term memory to capture contextual information of objects within the scene. Additionally, Lv et al. (2023) proposed a region-based adaptive association learning framework consisting of two sub-networks that extract features related to semantic distribution and contextual arrangement separately. Deep fusion networks are then employed to combine these features in a joint and boosting manner. However, in these studies, while different components for scene recognition are leveraged, the network backbones are still used as feature extractors.

Despite the progress made in scene recognition, the current methods primarily utilize deep neural networks as feature extractors, overlooking the unique characteristics of scenery images. Recent studies have shown promising results by introducing additional modules for scene recognition (Qiao et al., 2021; Yuan et al., 2022), exploring the use of adaptive learnable modules for assigning weights to different scales in scene recognition tasks. These approaches have demonstrated improved performance compared to conventional networks. However, the design of the network backbone, which forms the core architecture, has received limited attention in the existing literature. This reveals a gap between the state-of-the-art generic deep neural network design and the design of networks specifically tailored for scene recognition. It is crucial and urgent to incorporate the most advanced network design theories and practices into the field of scene recognition to bridge this gap and further advance the performance of scene recognition models.

3. Materials and method

The inception of the data-oriented network design can be traced back to our experimental observations. Specifically, we noted that modifying the depth and width of a network had varying impacts on object-centric and scene-centric data. To validate this hypothesis, we conducted a series of extensive experiments, which sought to confirm our initial conjecture.

Table 1
Top-1 and top-5 accuracy (%) comparison by changing the network depth.

Data	Model	Top-1	Top-5
Places365	ResNet-18	54.22	84.63
	ResNet-50	55.69	85.80
	ResNet-101	56.47	86.25
ImageNet	ResNet-18	70.52	89.56
	ResNet50	76.02	92.80
	ResNet-101	77.78	93.72

Table 2
Top-1 and top-5 accuracy (%) comparison by changing the network width. The numbers within the parenthesis are the width scaling factors.

Data	Model	Top-1	Top-5
Places365	ResNet-50 (×1)	55.69	85.80
	ResNet-50 (×2)	56.21	86.11
	ResNet-50 (×.5)	55.07	85.12
	ResNet-50 (×.25)	52.16	82.85
	ResNeXt-50	55.77	85.99
ImageNet	ResNet-50 (×1)	76.02	92.80
	ResNet-50 (×2)	78.51	94.09
	ResNet-50 (×.5)	72.08	90.78
	ResNet-50 (×.25)	64.04	85.76
	ResNeXt-50	77.80	94.30

3.1. Data sets and characteristics

3.1.1. Data sets

To understand the impact of network structure on datasets and ultimately applications, we use the ImageNet 2012 (Jia et al., 2009) and Places Standard dataset (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017b) as our evaluation datasets. ImageNet2012 and Places365 Standard are widely recognized as standard datasets that have significantly contributed to the advancement of computer vision research. ImageNet2012, introduced in 2012, comprises over 1.2 million meticulously labeled images across 1000 object categories, serving as a benchmark for training and evaluating object-centric image classification algorithms. An image in ImageNet 2012 usually contains a single object that is highly distinctive from the background. In contrast, Places365 Standard focuses on scene recognition, offering more than 1.8 million labeled images categorized into 365 diverse scene classes, including both indoor and outdoor scenes. The images in the Places365 Standard dataset present more complex scenery contents.

3.1.2. Impact of network structure to model performance

In deep learning, network structure and training data play important roles in model performance. Using ResNet as the baseline model, we conducted a comparison study using two popular datasets: Places365 and ImageNet. Table 1 presents the model performance by varying the network depth. Three ResNet variants are included: ResNet-18, ResNet-50, and ResNet-101. By increasing the network depth from 50 to 101, i.e., ResNet-50 and ResNet-101, we obtained a performance improvement of 1.40% and 2.32% on Place365 and ImageNet datasets, respectively, in terms of Top-1 accuracy. Theoretically, if widening the network is more effective in improving performance as stated in the literature, we are expecting more accuracy improvement when the network width is increased. Table 2 presents the model performance by varying network width. The network depth of all models is 50. Besides ResNet-50, ResNeXt-50 is included. By doubling the network width, we achieve an increase of top-1 accuracy by 3.28% using ImageNet, but only 0.94% using Place365. More surprisingly, for ResNeXt, which also doubled the network width, the relative performance increase on ImageNet is 2.34% in terms of top-1 accuracy, but the number is only 0.14% on Place365.

The trend that model performance on ImageNet is more sensitive to changes in network width compared to Place365 is also observed when we decrease network depth or narrow down the width. When the network depth was reduced from 50 to 18, ImageNet experienced a 7.23% relative top-1 accuracy decrease, while for Place365 it was 2.71%. The drop of top-1 performance using ImageNet was approximately 2.7 times than that of the top-1 performance drop using Places. However, when we reduce the width of ResNet-50 by half, this ratio changed to 4.7. The claim that widening the network might provide a more effective way to improve performance is probably biased towards ImageNet (an object-centric dataset) and overlooks the characteristics of scenery images.

Table 3
Top-1 and top-5 accuracy (%) comparison by changing the network depth using 100 randomly selected classes.

Data	Model	Top-1	Top-5
Places365	ResNet-18	70.63	94.32
	ResNet-50	71.14	94.56
ImageNet	ResNet-18	81.17	94.50
	ResNet50	83.19	95.42

Table 4
Top-1 and top-5 accuracy (%) comparison by changing the network width using 100 randomly selected classes. The numbers within the parenthesis are the width scaling factors.

Data	Model	Top-1	Top-5
Places365	ResNet-50 (×1)	71.14	94.56
	ResNet-50 (×.5)	71.04	94.21
	ResNet-50 (×.25)	70.14	93.76
ImageNet	ResNet-50 (×1)	83.19	95.42
	ResNet-50 (×.5)	80.17	94.06
	ResNet-50 (×.25)	76.93	92.62

Since Places365 has 365 classes while ImageNet has 1000 classes, to avoid the bias caused by the different number of classes between the two datasets, we randomly selected 100 classes from each dataset and conducted comparison experiments. As we reduced the size of the datasets, we observed under-fitting when we applied relatively small datasets to large models. Therefore, we only include the comparison results for small models in Tables 3 and 4. We observed that for the Places365 dataset, switching from ResNet-50 to ResNet-18 led to a 0.72% top-1 accuracy drop on Places and a 2.49% top-1 accuracy drop on ImageNet, i.e., the performance drop on ImageNet was around 3.46 times the drop on Places. Meanwhile, narrowing the width of ResNet-50 to 1/2 and 1/4 of the original width led to a 0.14% and 1.41% top-1 performance drop on Places365. For ImageNet, the accuracy decrease was 3.6% and 7.5%. The performance drop caused by halving the width on ImageNet was 25.7 times the performance drop on Places, which verified our hypothesis that altering network width has a less significant effect on scenery data and is not biased towards the number of classes.

3.1.3. Complexity of images

Our first hypothesis is that the performance difference is caused by the complexity of scenery images. This hypothesis is based on the distinct complexity difference between scene images and object-centric images: object-centric images contain one dominant object that occupies a large portion of the view, while scene images consist of multiple objects in different sizes and background clutters. Fig. 1 shows two samples from the benchmark object-centric dataset (ImageNet) and the benchmark scene dataset (Place365), respectively. Fig. 1(a) is labeled as “bald eagle”, in which the eagle stands in the center of the view and occupies a large portion of the entire image. Fig. 1(b) is labeled as “forest-broad leaf”, and the entire view consists of not only a bird but a few tree branches and leaves. As the correct recognition of scenery images relies on multiple components, a scene image is typically considered more complex than an object-centric image.

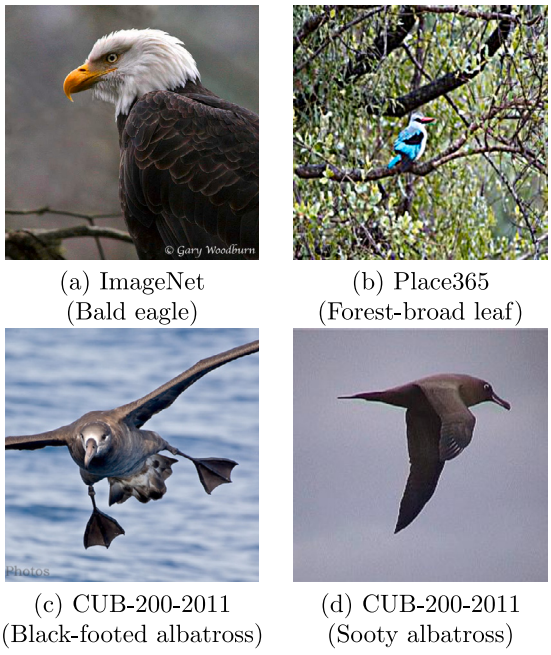


Fig. 1. Example images from benchmark object recognition dataset ImageNet (a), benchmark scene recognition dataset Places365 (b), and fine-grained dataset Caltech-UCSD Birds-200-2011 (c and d).

To evaluate this hypothesis, we use CUB-200-2011 dataset that is widely known to be "complex" fine-grained image classification dataset. The fine-grained classification is considered a more complex task as the classes in the dataset can only be discriminated by local and subtle differences. CUB-200-2011 consists of 200 different species of birds, which serves as a benchmark dataset for fine-grained classification tasks. Fig. 1(c) and (d) depict samples of the CUB-200-2011 dataset. In Fig. 1, the black-footed albatross (c) and sooty albatross (d) are considered two different categories in classification. The two albatrosses are similar in appearance, and differentiating them is challenging due to the subtle traits that characterize the different species.

We conducted experiments using the two complex datasets (Places365 and CUB-200-2011) and one object centric dataset (ImageNet). The results are shown in Table 5. Using the benchmark ResNet-50 as the backbone, we observed that on CUB-200-2011, the relative top-1 accuracy increased by 1.81% when doubling the width and dropped by 4.28% when we narrowed the width to half of the original. This performance change caused by altering the width is much more acute compared to the result on Places365 (0.94% and 1.11%, respectively, for doubling and halving the width) under the same settings, which demonstrated that a wide network is able to effectively enhance the recognition of complex, fine-grained features.

Table 5
Top-1 accuracy (%) of ResNet-50 on Places365, ImageNet, and Caltech-UCSD Birds-200–2011 by changing network width.

Width Scaling Factor	Places	CUB	ImageNet
2	56.21	71.53	78.51
1	55.69	70.26	76.02
0.5	55.07	67.25	72.08
0.25	52.16	61.29	64.04

This result does not support our first hypothesis: if the performance difference originated from the complexity of the data, we should observe a moderate performance change on the CUB-200-2011 dataset along with the changing of network width. The results demonstrate that the performance variance on different datasets is not a consequence of the complexity (rich fine-grained details) of the data. We found that the variance in performance across different datasets cannot be solely attributed to the level of complexity, specifically the rich fine-grained details.

3.2. Deep-narrow network for scenery image classification

3.2.1. Spatial and channel features for scenery image recognition

Based on our observations, we formulated a second hypothesis: for scene recognition tasks, learning spatial information is more crucial than learning more fine-grained features. As defined in Fan, Xian, Losch, and Schiele (2020), spatial information refers to the spatial ordering on the feature map. Intuitively, the semantic meaning related to spatial layout is limited for images that only contain one object, but for scene images, the spatial structures or contextual information are likely to contribute more to the understanding of the scene. Therefore, we hypothesized that learning spatial information is more important for scene recognition tasks.

We conducted the experiments by gradually feeding low and high-frequency information of Place365 and ImageNet datasets to the network. Generally speaking, the high-frequency information in the image refers to the regions where the intensity of the image (brightness/gray-scale) changes drastically, which are often called the edges or boundaries; the low-frequency information in the image refers to the regions where the image intensity changes smoothly, such as large patches of color. As shown in Fig. 2, the image filtered by low-pass filters tends to present proximate or blurred patterns/features of the original image, the images filtered by high-pass filters better preserved the spatial information, i.e., high-frequency information are more spatially informative features compared to low-frequency information.

To understand the importance of low and high-frequency information in different datasets, we designed low-pass and high-pass filters based on Fourier Transform. We transformed the testing images into the frequency domain using Fourier Transform and applied both low and high-pass filters to test how low/high-frequency information can affect the model performance on different datasets. Fig. 3 shows the design of the filters: for low-pass filters, we masked the high-frequency

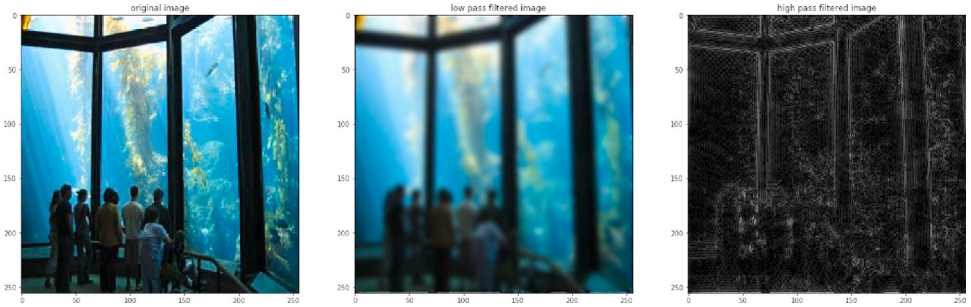


Fig. 2. Results on scenery image using low-pass and high-pass filters.

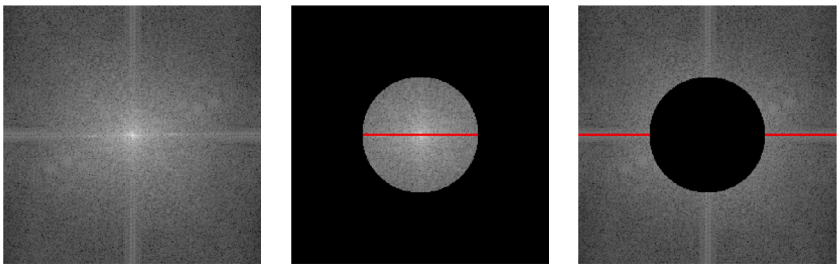


Fig. 3. Illustration of low and high pass filters. In (b) and (c), the mask (black region) denotes information removed by corresponding filters; the length of the red lines denotes the corresponding filter size.

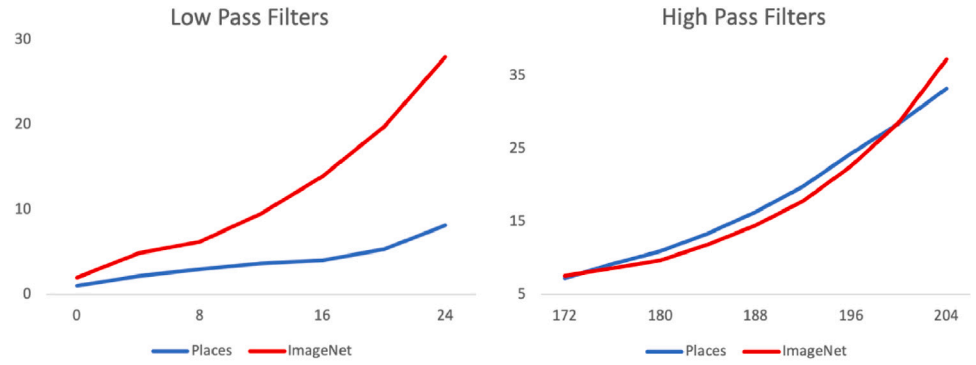


Fig. 4. Top-1 accuracy (%) on Place365 and ImageNet datasets using low-pass filters (left) and high-pass filters (right). The x-axis denotes the size of the corresponding filters. Note that for the 100-class ImageNet and Place365 data sets, the top-1 accuracy on ImageNet is 16.9% higher.

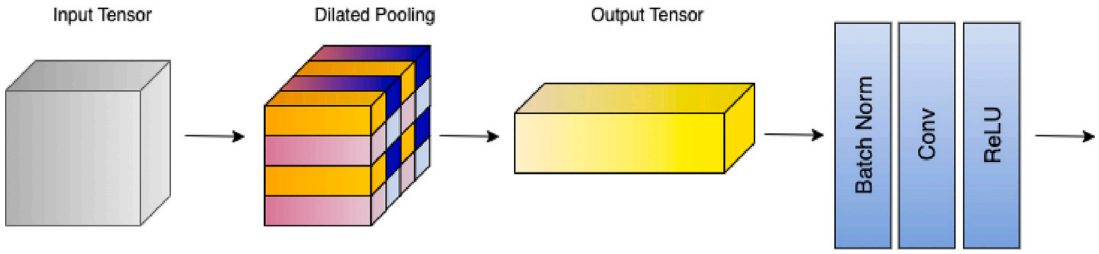


Fig. 5. In the schema of a Deep-Narrow Network, we increased depth and decreased the width of the network and added a Dilated Pooling module to persevere more spatial information.

components, and for high-pass filters, we masked the low-frequency components. Low pass filters preserve low-frequency features; high pass filters preserve high-frequency/spatial structure information.

For a fair comparison, we randomly selected 100 classes from the Place365 and ImageNet datasets to conduct the experiments. Note that scene recognition is considered a harder task compared to object recognition, so the classification accuracy on Place365 is lower despite the chance being the same. The results are shown in Fig. 4. In both sub-figures, the x-axis represents the size of the corresponding low/high pass filter in the spectrum domain (with a maximum size of 224), and the y-axis denotes the top-1 accuracy. Through comparison, we observed some enlightening phenomena: when gradually feeding the low-frequency information to the networks, the performance increase on ImageNet is steeper than on Place365 (Fig. 4(a)). Surprisingly, when using a low-pass filter of size 33, ImageNet achieved a top-1 accuracy of nearly 30%, while the chance is only 1%. This suggests that correct recognition of object-centric images heavily relies on low-frequency information.

On the contrary, when gradually feeding the high-frequency information to the networks, we observe that the model trained on scenery datasets is more sensitive to high-frequency information (Fig. 4(b)). Notably, when the size of the high-pass filter is around 210 to 214, the top-1 accuracy on Place365 exceeds the top-1 accuracy on ImageNet, despite the classification accuracy on ImageNet is 16.9% higher than on Place365 for 100 classes. This observation demonstrates that recognition of scenery images is susceptible to high-frequency information.

Our observations perfectly fit the experimental results above. The high and low-frequency information in images approximately represent the learned spatial and channel-wise information in deep networks. Wide networks have an expanded number of channels, which enables the network to learn more fine-grained features. Deep networks have an increasing number of layers and larger receptive fields, which enables the network to learn more spatial and scale information. Our experimental results support the hypothesis that learning spatial information is more crucial than fine-grained features for scene recognition tasks.

3.2.2. *Deep-narrow structure*

Based on our observations, we argue that designing networks with a larger depth and a smaller width can potentially be an effective and efficient option for correctly recognizing scenery images, as this task is highly dependent on learning spatial information. As shown in Figs. 5

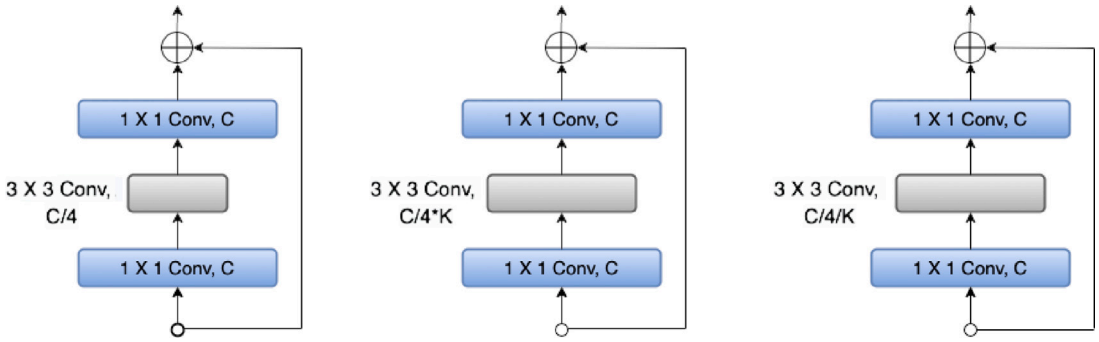
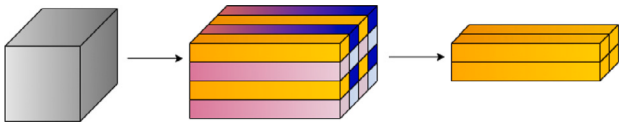
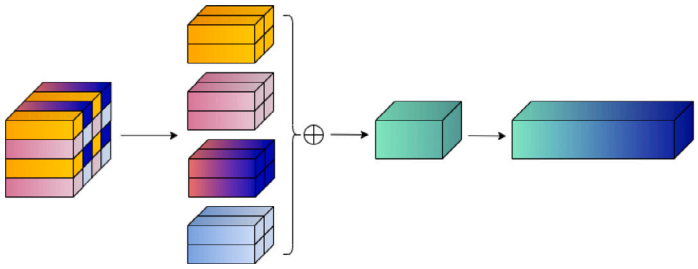


Fig. 6. The convolution block of ResNet (left), Wide ResNet (middle), and Deep-Narrow Network (right). C denotes the number of channels.



(a) Down-sample component in ResNet, during the down-sampling process, 3/4 of the spatial information is discarded and the number of channels is quadrupled.



(b) Dilated Pooling module which does not discard spatial information.

Fig. 7. The schema of the down-sample component in ResNet and Dilated Pooling module.

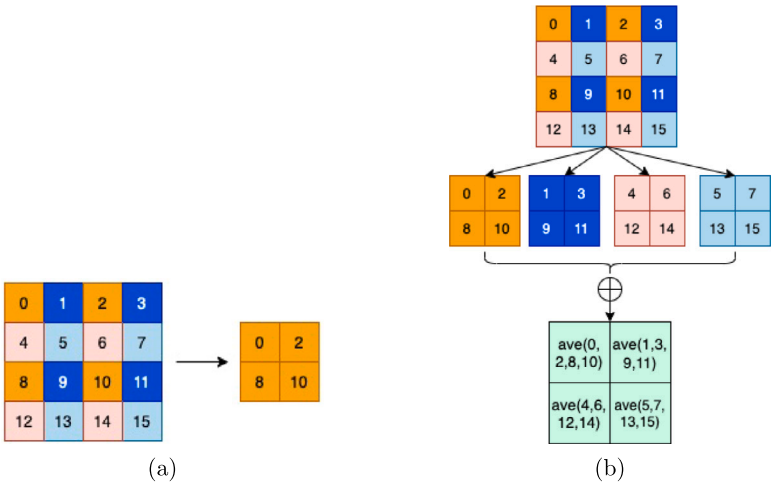


Fig. 8. The schema of the down-sample component in ResNet and Dilated Pooling module from the view of spatial dimension. (a) The down-sample component in ResNet from the view of spatial dimension. During the down-sampling process, 3/4 of the spatial information is discarded and the number of channels is doubled. (b) Dilated Pooling module from the view of spatial dimension.

and 6, we propose a Deep-Narrow architecture that differs from the popular Wide ResNet (Zagoruyko & Komodakis, 2016), which uses the factor K to increase the width of the ResNet. Instead, we increase the number of layers in ResNet and decrease the width of the network using width factor K. In our designed Deep-Narrow Network, the value of K is set to 2, i.e., we halve the network width used in benchmark ResNet.

3.2.3. *Dilated pooling*

In addition to deepening the network to better process spatial information, we have also designed a Dilated Pooling module to better preserve the spatial information in ResNet. As shown in Fig. 7(a), in ResNet, the downsampling process involves quadrupling the width of the network and discarding 3/4 of the features along the spatial

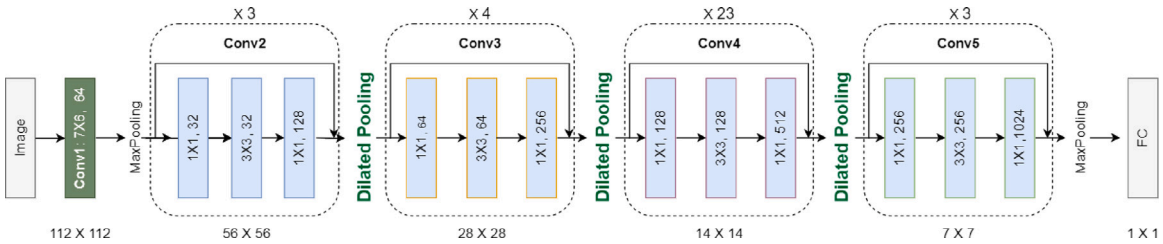


Fig. 9. The schema of and Deep-Narrow Network with Dilated Pooling. The numbers inside convolution blocks denote the kernel size and number of channels in each block. The numbers under the figure stand for the spatial dimension size of the corresponding feature maps.

Table 6
GFLOPS and number of Parameters using ResNet with different designs and Dilated Pooling (DP) module. The number of parameters is in million. Please note that for the cases using Places365-Standard dataset, we calculate the number of parameters based on 365-class models (Place365).

Data	Model	GFLOPs	# Params
Places365	ResNet-50	4.12	24.26
	Deep-Narrow Network	2.00	11.03
	Deep-Narrow (with DP)	2.00	11.03
ImageNet	ResNet-50	4.12	25.56
	Deep-Narrow Network	2.00	11.68
	Deep-Narrow (with DP)	2.00	11.68

Table 7
Top-1 and Top-5 accuracy (%) using ResNet with different depth and width. The numbers within the parenthesis are the width scaling factors.

Data	Model	Top-1	Top-5
Place365	ResNet-50 (×1)	55.69	85.80
	ResNet-50 (×.5)	55.07	85.12
	Deep-Narrow Network	55.58	85.80
ImageNet	ResNet-50 (×1)	76.02	92.80
	ResNet-50 (×.5)	72.08	90.78
	Deep-Narrow Network	74.99	92.31

Table 8
Top-1 and top-5 accuracy (%) using Deep-Narrow Network and Dilated Pooling (DP) module. The numbers within the parenthesis are the width scaling factors. The best results are highlighted in bold.

Data	Model	Top-1	top-5
Place365	ResNet-50 (×1)	55.69	85.80
	Deep-Narrow	55.58	85.80
	Deep-Narrow (with DP)	**55.91**	**86.12**
ImageNet	ResNet-50 (×1)	76.02	92.80
	Deep-Narrow	74.99	92.31
	Deep-Narrow (with DP)	74.63	92.13

dimension. Based on our findings, this design is not suitable for scene recognition as discarding spatial information is likely to significantly impact the performance. Alternatively, we designed a Dilated Pooling module to better preserve the spatial information (Fig. 7(b)).

Further details are illustrated in Fig. 8. In a Dilated Pooling module, instead of directly discarding 3/4 of the spatial information, we divide the feature maps into four sub-sections along the spatial dimension. We then conduct convolution on the four feature maps and merge the results together via summation operation. By leveraging Dilated Pooling, we can use all the spatial information without increasing the number of FLOPs and parameters.

Fig. 9 illustrates the overall network architecture of our proposed method. This network extends ResNet by integrating our Deep-Narrow structure and dilated pooling modules. The dilated pooling modules are embedded in between the convolution blocks, which are implemented with the Deep-Narrow structures.

4. Experimental results and discussion

4.1. Experiment settings

We train deep network models and compute single-crop (224 × 224 pixels) top-1 and top-5 accuracy based on the application of the models to the validation set. We train each model for 100 epochs on eight Tesla V100 GPUs with 32 images per GPU (the batch size is 256). All models are trained using synchronous SGD (Stochastic Gradient Descent) with a Nesterov momentum of 0.9 and a weight decay of 0.0001. The learning rate is 0.1 and is reduced by a factor of 10 in every 30 epochs. For training ResNet and its variants, we follow the settings in He et al. (2016).

4.2. Computational cost analysis

This section presents an evaluation of the computational efficiency of our proposed design compared to the canonical ResNet-50. Table 6 shows the results of our evaluation in terms of computation efficiency, measured in GFLOPs, and the number of parameters in millions.

Our design achieved a significant improvement in computational efficiency, with only 2.00 GFLOPs consumed, which is less than half of the 4.12 GFLOPs consumed by the canonical ResNet-50. Additionally, we were able to reduce the number of parameters used in our design to 11.68 million, compared to the 24.26 million parameters used in ResNet-50. Furthermore, we evaluated the impact of inserting our Dilated Pooling (DP) module on the efficiency of the network. Our results show that the efficiency of the network is not compromised when the Dilated Pooling module is inserted. We will further show the reduction in parameters is achieved without sacrificing the performance of the network. With the help of DP module, we can even improve the model accuracy on scenery data set.

4.3. Performance analysis

Table 7 shows the performance comparisons among the benchmark ResNet-50 (i.e., ResNet-50 (× 1)), ResNet-50 with half the width (i.e., ResNet-50 (×.5)), and our Deep-Narrow Network on the Place365 dataset. The Deep-Narrow Network achieves comparable evaluation scores with the benchmark ResNet-50 while using less than half of the FLOPs and parameters, with a relative top-1 accuracy drop of only 0.20%. On the ImageNet dataset, the Deep-Narrow Architecture obtained a relative top-1 accuracy drop of 1.35%. These results reaffirm that scene recognition is highly dependent on learning spatial information. We demonstrate that data have their preferences and the network design should rely on the characteristics of the data rather than blindly following the assertion that widening the network might provide a more effective way to improve model performance over making ResNet deeper.

As shown in Table 8, we integrated the Dilated Pooling module with our Deep-Narrow Network design, resulting in better performance than

Table 9
GFLOPS and number of Parameters using ResNet with different designs using different backbones. The numbers within the parenthesis are the width scaling factors. The number of parameters is in million. Note that we calculate the number of parameters based on 365-class models (Place365). The best results are highlighted in bold.

Model	GFLOPs	# Params
ResNet-50 (×1)	4.12	24.26
ResNet-Ave	2.26	11.03
ResNet-Max	2.26	11.03
ResNet-D	2.26	11.03
Antialiased-CNN	2.26	11.03
Deep-Narrow Network	**2.00**	**11.03**
Deep-Narrow Network (with DP)	**2.00**	**11.03**

Table 10
Top-1 and top-5 accuracy (%), precision, recall and F-1 score on Place365 using different backbones. The numbers within the parenthesis are the width scaling factors. The best results are highlighted in bold, second best results are underlined.

Model	Top-1	Top-5	Precision	Recall	F-1
ResNet-50 (×1)	55.69	85.80	56.34	54.76	55.53
ResNet-Ave	55.60	85.58	56.51	54.95	55.71
ResNet-Max	55.55	85.56	56.16	54.53	55.33
ResNet-D	55.79	<u>85.93</u>	56.45	54.87	55.65
Antialiased-CNN	<u>55.85</u>	<u>85.93</u>	**56.92**	<u>55.13</u>	<u>56.01</u>
Deep-Narrow Network	55.58	85.80	56.13	54.53	55.32
Deep-Narrow Network (with DP)	**55.91**	**86.12**	<u>56.84</u>	**55.26**	**56.04**

the benchmark ResNet on the Place365 dataset while using less than half of the FLOPs and the number of parameters. Specifically, adding Dilated Pooling to our Deep-Narrow Network resulted in a relative top-1 accuracy increase of 0.59% (Table 8), while causing only a 0.48% relative top-1 accuracy drop on ImageNet. This demonstrates the effectiveness and efficiency of our data-oriented network design approach and emphasizes the importance of designing networks according to the characteristics of the data.

4.4. Comparison with the state-of-the-art methods

To evaluate the effectiveness of our proposed method, we compared it against several state-of-the-art approaches that aimed to minimize information loss caused by reducing spatial resolution. These approaches included ResNet-D (He et al., 2019) and Antialiased-CNN (Zhang, 2019), which employed average pooling and convolution-based pooling strategies, respectively, to preserve more spatial information. Additionally, we implemented two baseline strategies, ResNet-Ave and ResNet-Max, which utilized averaging and max-pooling operations instead of discarding three-quarters of the information.

Table 9 presents the resource efficiency of our model compared to the aforementioned methods. Overall, our method used 0.26 fewer GLOPs compared to Antialiased-CNN and variants of ResNets, representing an 11.5% computational resource reduction. In contrast to ResNet-50, i.e., our baseline method, the computation reduction is by more than 50%. This advantage is also supported by the number of parameters. Note that the unit for the number of parameters is millions. The size of our proposed network is only 45.5% of that of the ResNet-50.

Table 10 presents a comprehensive comparison of different models based on their performance metrics for scene recognition using the Place365 dataset. Our Deep-Narrow Network with Dilated Pooling achieves the highest top-1 accuracy of 55.91% among all the models. Moreover, it attains the highest top-5 accuracy of 86.12%, outperforming the other models. These results demonstrate the model's ability to provide accurate predictions within the top 5 classes for the given scenes. Furthermore, the Deep-Narrow Network with Dilated Pooling consistently achieves greater precision, recall, and F-1 scores. In summary, our results indicate that the Deep-Narrow Network with Dilated Pooling achieves superior accuracy compared to the state-of-the-art methods while utilizing much fewer computational resources.

4.5. Case analysis

Fig. 10 depicts instances where the vanilla ResNet model misclassifies the images, while our proposed Deep-Narrow Network with Dilated

Fig. 10. Examples that are miss-classified by the vanilla ResNet but correctly classified by our method.

Fig. 11. Examples that are miss-classified by our method.

pooling accurately classifies them. These images represent diverse indoor and outdoor scenarios, exhibiting intricate spatial information. The successful recognition of these scenes by the Deep-Narrow Network with Dilated Pooling underscores its ability to effectively handle scenes with intricate spatial details.

Fig. 11 illustrates examples of misclassified images from our experiments. We observed that the misclassification can be categorized into three distinct types: (1) misclassification of object-centric images. When an image contains a prominent object, such as the airplane in Fig. 11(a), with a limited or unrepresentative background, our model may focus too much on the object and misclassify the overall scene. (2) misclassification of scene mixing images. These images combine different scenes, leading to confusion for the classifier. In Fig. 11(b), the amusement park includes a building resembling a castle, which creates ambiguity and challenges the classifier's ability to assign the correct scene category accurately. (3) misclassification of rare cases. The background of an image can be misleading. In Fig. 11(c), the apartment buildings situated near a river, a scenario not well-represented in our training dataset, can lead the classifier to mistakenly associate it with a river house scene.

To examine how our models preserve detailed spatial information, we utilized Grad-CAM (Selvaraju et al., 2017) as a visualization tool. Fig. 12 presents the heat map of six scene images using different network architectures. The row of "ImageNet" shows the results of a ResNet trained with the ImageNet dataset, while the row of "Places365" shows the results of a ResNet trained with the Places365 dataset. By comparing these heat maps, we can evaluate the preservation of spatial details in our proposed models. Fig. 12 depicts that our method tends to preserve more spatial details, which are indicative of the semantic meaning of the scene. For example, when examining images of categories "building facade", "castles", and "water tower", our model successfully captures distinguishing architectural elements such as windows, minibars in mosques, and water tanks with their bases. Additionally, our model not only captures prominent components like flowers in a greenhouse, furnishings in a banquet, and desks in office cubicles but also identifies specific features like greenhouse shelves, tableware, and computers. These additional details serve as strong indicators of the respective scenes. The ability to leverage these detailed spatial features empowers our Deep-Narrow Network with Dilated Pooling to achieve favorable performance on the scene recognition task while utilizing reduced computational resources. This highlights the effectiveness of our model in capturing and utilizing spatial information.

5. Conclusion

This paper investigates the impact of complex scenery images on network architecture design using ImageNet (an object recognition

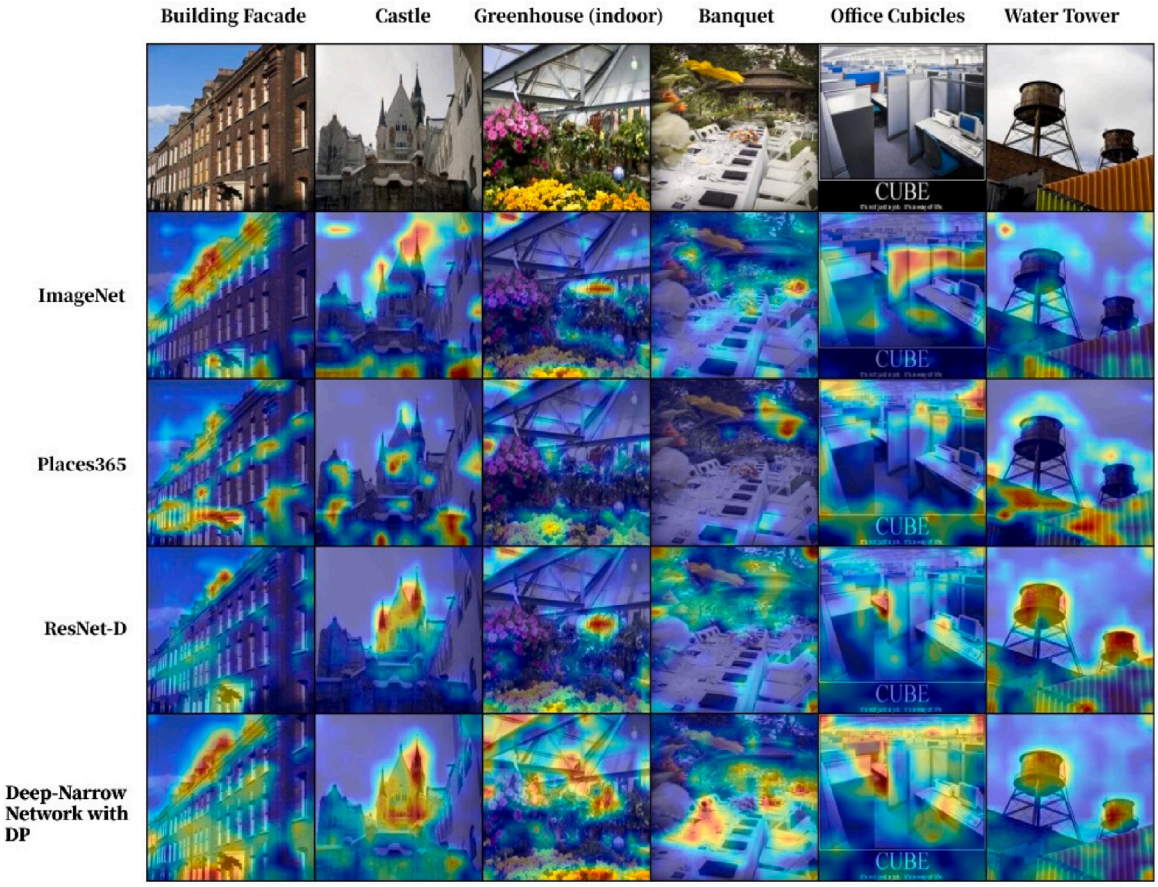


Fig. 12. Heat maps of various images generated using Grad-CAM (Selvaraju et al., 2017). Warmer color denotes a higher value. "ImageNet" and "Place365 Standard" denotes vanilla ResNet pre-trained using the corresponding dataset. Our design preserves comprehensive information that represents the semantic meaning of the scenes.

dataset) and Places365 (a scene recognition dataset) as examples. Through a series of carefully designed experiments, we demonstrate that the characteristics of datasets can significantly affect the performance of models. Specifically, wider networks tend to perform better in recognizing images with prominent objects but have less impact on recognizing scenery images. We further validated this hypothesis through comparison experiments and showed that learning spatial information is more critical in scene recognition tasks compared to object classification tasks. This explains why deepening the networks is more effective than widening them for scene recognition, as deeper networks can better learn spatial information in the training examples. Therefore, deploying networks with a greater depth and a smaller width and emphasizing spatial information learning can benefit scene recognition backbone designs. Our proposed Deep-Narrow Network and Dilated Pooling module demonstrate the effectiveness and efficiency of taking advantage of data properly. Our design achieves better accuracy than the benchmark ResNet-50 on the canonical scenery data set using less than half of the computation resources.

Scene recognition continues to pose challenges, primarily due to limitations in available datasets and gaps in theoretical research. In our future work, we plan to investigate self-supervised learning methods. The recent integration of self-supervised learning techniques shows promise in reducing the dependence on labeled scenery data. By leveraging unlabeled samples, meaningful representations can be learned, thereby enhancing scene recognition capabilities, and enabling better real-world applications. In addition, taking advantage of recent advancements in computer vision foundational models and artificial intelligence-driven graphics computing presents new opportunities for advancing scene recognition.

CRediT authorship contribution statement

Zhinan Qiao: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Xiaohui Yuan:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **Runmei Zhang:** Software, Writing – review & editing. **Tian Chen:** Resources, Formal analysis, Writing – review & editing. **Chaoning Zhang:** Resources, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Bordelon, B., & Pehlevan, C. (2022). Self-consistent dynamical field theory of kernel evolution in wide neural networks. *NeurIPS*, *35*, 32240–32256.

Chen, Y., Luo, Z., Li, W., Lin, H., Nurunnabi, A., Lin, Y., et al. (2022). WGNet: Wider graph convolution networks for 3D point cloud classification with local dilated connecting and context-aware. *International Journal of Applied Earth Observation and Geoinformation*, *110*, Article 102786.

Cheng, X., Lu, J., Feng, J., Yuan, B., & Zhou, J. (2018). Scene recognition with objectness. *Pattern Recognition*, *74*, 474–487.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *CVPR* (pp. 1251–1258).

Fan, Y., Xian, Y., Losch, M. M., & Schiele, B. (2020). Analyzing the dependency of convnets on spatial information. In *DAGM German conference on pattern recognition* (pp. 101–115). Springer.

Guo, S., Wang, Y., Li, Q., & Yan, J. (2020). Dmcp: Differentiable markov channel pruning for neural networks. In *CVPR* (pp. 1539–1547).

Gupta, S., Sharma, K., Dinesh, D. A., & Thenkanidiyoor, V. (2021). Visual semantic-based representation learning using deep CNNs for scene recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, *17*(2), 1–24.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2019). Bag of tricks for image classification with convolutional neural networks. In *CVPR* (pp. 558–567).

Herranz, L., Jiang, S., & Li, X. (2016). Scene recognition with CNNs: objects, scales and dataset bias. In *CVPR* (pp. 571–579).

Jia, D., Wei, D., Richard, S., Li-Jia, L., Kai, L., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *NeurIPS*, *25*, 1097–1105.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2014). The CIFAR-10 dataset. URL https://www.cs.toronto.edu/~kriz/cifar.html.

Li, J., Hassani, A., Walton, S., & Shi, H. (2023). ConvMLP: Hierarchical convolutional mlps for vision. In *CVPR* (pp. 6306–6315).

Lin, C., Lee, F., Xie, L., Cai, J., Chen, H., Liu, L., et al. (2022). Scene recognition using multiple representation network. *Applied Soft Computing*, *118*, Article 108530.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *ECCV* (pp. 740–755). Springer.

Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width. In *NeurIPS* (pp. 6232–6240).

Lv, G., Dong, L., Zhang, W., & Xu, W. (2023). Region-based adaptive association learning for robust image scene recognition. *The Visual Computer*, *39*(4), 1629–1649.

Mirzadeh, S. I., Chaudhry, A., Yin, D., Hu, H., Pascanu, R., Gorur, D., et al. (2022). Wide neural networks forget less catastrophically. In *ICML* (pp. 15699–15717). PMLR.

Nguyen, Q., & Hein, M. (2018). Optimization landscape and expressivity of deep CNNs. In *International conference on machine learning* (pp. 3730–3739). PMLR.

Nguyen, T., Raghu, M., & Kornblith, S. (2021). Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. In *ICLR*. Vienna, Austria.

Qiao, Z., Yuan, X., & Elhoseny, M. (2020). Urban scene recognition via deep network integration. In *International conference on urban intelligence and applications* (pp. 135–149). Springer.

Qiao, Z., Yuan, X., Zhuang, C., & Meyarian, A. (2021). Attention pyramid module for scene recognition. In *ICPR* (pp. 7521–7528). Milan, Italy.

Radhakrishnan, A., Belkin, M., & Uhler, C. (2023). Wide and deep neural networks achieve consistency for classification. *Proceedings of the National Academy of Sciences*, *120*(14), Article e2208779120.

Rehman, A., Saleem, S., Khan, M. U., Jabeen, S., & Shafiq, M. (2021). Scene recognition by joint learning of DNN from bag of visual words and convolutional DCT features. *Applied Artificial Intelligence*, *35*, 1–19.

Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Visual explanations from deep networks via gradient-based localization. In *ICCV* (pp. 618–626).

Seong, H., Hyun, J., Chang, H., Lee, S., Woo, S., & Kim, E. (2019). Scene recognition via object-to-scene class conversion: end-to-end training. In *IJCNN* (pp. 1–6).

Seong, H., Hyun, J., & Kim, E. (2020). Fosnet: An end-to-end trainable deep neural network for scene recognition. *IEEE Access*, *8*, 82066–82077.

Shen, X., Wang, Y., Lin, M., Huang, Y., Tang, H., Sun, X., et al. (2023). DeepMAD: Mathematical architecture design for deep convolutional neural network. In *CVPR* (pp. 6163–6173).

Shi, J., Zhu, H., Yu, S., Wu, W., & Shi, H. (2019). Scene categorization model using deep visually sensitive features. *IEEE Access*, *7*, 45230–45239.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR*. San Diego, USA.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. In *AAAI* (pp. 4278–4284).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *CVPR* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR* (pp. 2818–2826).

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML* (pp. 6105–6114). PMLR.

Veit, A., Wilber, M., & Belongie, S. (2016). Residual networks are exponential ensembles of relatively shallow networks. In *NeurIPS*. Barcelona, Spain.

Wang, C., Peng, G., & De Baets, B. (2020). Deep feature fusion through adaptive discriminative metric learning for scene recognition. *Information Fusion*, *63*, 1–12.

Wang, Z., Wang, L., Wang, Y., Zhang, B., & Qiao, Y. (2017). Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *IEEE Transactions on Image Processing*, *26*(4), 2028–2041.

Xia, S., Zeng, J., Leng, L., & Fu, X. (2019). WS-AM: Weakly supervised attention map for scene recognition. *Electronics*, *8*(10), 1072.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *CVPR* (pp. 1492–1500).

Yuan, X., Qiao, Z., & Meyarian, A. (2022). Scale attentive network for scene recognition. *Neurocomputing*, *492*, 612–623.

Zagoruyko, S., & Komodakis, N. (2016). Wide residual networks. In *BMVC*. City of York, United Kingdom.

Zhang, R. (2019). Making convolutional networks shift-invariant again. In *International conference on machine learning* (pp. 7324–7334). PMLR.

Zhang, C., Benz, P., Argaw, D. M., Lee, S., Kim, J., Rameau, F., et al. (2021). Resnet or densenet? Introducing dense shortcuts to resnet. In *WACV* (pp. 3550–3559).

Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., et al. (2022). Resnest: Split-attention networks. In *CVPR* (pp. 2736–2746).

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017a). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017b). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464.
Zhu, Y., Deng, X., & Newsam, S. (2019). Fine-grained land use classification at the city scale using ground-level images. *IEEE Transactions on Multimedia*, *21*(7), 1825–1838.
Zoph, B., & Le, Q. V. (2017). Neural architecture search with reinforcement learning. In *ICLR*. Toulon, France.