Paleobiology, 49(1), 2023, pp. 1–22
DOI: 10.1017/pab.2022.14

PALEOBIOLOGY
A PUBLICATION OF THE Paleontological SOCIETY

Featured Article

Automatic taxonomic identification based on the Fossil Image Dataset (>415,000 images) and deep convolutional neural networks

*Xiaokang Liu, Shouyi Jiang, Rui Wu, Wenchao Shu, Jie Hou, Yongfang Sun, Jiarui Sun, Daoliang Chu, Yuyang Wu, and Haijun Song**

Abstract.—The rapid and accurate taxonomic identification of fossils is of great significance in paleontology, biostratigraphy, and other fields. However, taxonomic identification is often labor-intensive and tedious, and the requisition of extensive prior knowledge about a taxonomic group also requires long-term training. Moreover, identification results are often inconsistent across researchers and communities. Accordingly, in this study, we used deep learning to support taxonomic identification. We used web crawlers to collect the Fossil Image Dataset (FID) via the Internet, obtaining 415,339 images belonging to 50 fossil clades. Then we trained three powerful convolutional neural networks on a high-performance workstation. The Inception-ResNet-v2 architecture achieved an average accuracy of 0.90 in the test dataset when transfer learning was applied. The clades of microfossils and vertebrate fossils exhibited the highest identification accuracies of 0.95 and 0.90, respectively. In contrast, clades of sponges, bryozoans, and trace fossils with various morphologies or with few samples in the dataset exhibited a performance below 0.80. Visual explanation methods further highlighted the discrepancies among different fossil clades and suggested similarities between the identifications made by machine classifiers and taxonomists. Collecting large paleontological datasets from various sources, such as the literature, digitization of dark data, citizen-science data, and public data from the Internet may further enhance deep learning methods and their adoption. Such developments will also possibly lead to image-based systematic taxonomy to be replaced by machine-aided classification in the future. Pioneering studies can include microfossils and some invertebrate fossils. To contribute to this development, we deployed our model on a server for public access at www.ai-fossil.com.

Xiaokang Liu, Shouyi Jiang, Rui Wu, Wenchao Shu, Jie Hou, Yongfang Sun, Jiarui Sun, Daoliang Chu, Yuyang Wu, and Haijun Song. State Key Laboratory of Biogeology and Environmental Geology, School of Earth Sciences, China University of Geosciences, Wuhan 430074, China. E-mail: xkliu@cug.edu.cn, jsy@cug.edu.cn, ruicug@163.com, wenchaoshu@live.cn, hjbb@cug.edu.cn, yongfangsun@cug.edu.cn, s@cug.edu.cn, chudl@cug.edu.cn, wuyuyang01@gmail.com, haijunsong@cug.edu.cn

Accepted: 28 March 2022
*Corresponding author.

Introduction

Systematic paleontology is essential work in paleontology and biostratigraphy, because it helps reveal biological evolution in deep time. Despite the accurate taxonomic identification of fossils being essential for research, traditional methods rely only on individual identification, which is time-consuming, labor-intensive, and subjective. Alternatively, machine learning can help experts identify biological specimens and significantly increase efficiency (MacLeod et al. 2010), as was intended since it was first proposed 50 years ago (Pankhurst 1974). Traditional machine learning methods rely on a set of characteristics (e.g., geometric features) selected or designed manually by experts. Analytical methods can benefit from the use of machine learning methods, such as shallow artificial neural networks, support vector machines, decision trees, clustering, and naive Bayes classifiers, to handle nonlinear complex tasks (MacLeod 2007), achieving a suitable efficiency on small datasets. Owing to the development of computer science and the advent of big data, deep learning has advanced substantially over the past decade (LeCun et al. 2015), enabling the analysis of massive, high-dimensional, and complex data

0094-8373/23

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

(Hinton and Salakhutdinov 2006). Remarkably, in the field of computer vision, deep convolutional neural networks (DCNNs) (Krizhevsky et al. 2012; Szegedy et al. 2015) can learn and automatically extract features from images. In the past few years, convolutional neural networks (CNNs) have been increasingly applied in geoscience and its multidisciplinary fields, as reported in the Web of Science (Fig. 1). Meanwhile, previous studies have proven the feasibility of using deep learning methods for the automatic identification of biotics or fossils (MacLeod et al. 2010; Romero et al. 2020).

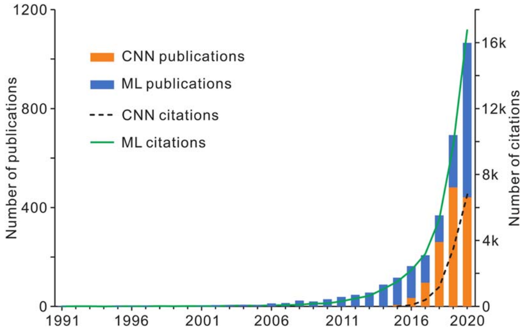

FIGURE 1. Statistics of publications and citations of the topics of machine learning (ML) and convolutional neural networks (CNNs) in geoscience and its multidisciplinary fields from Web of Science (to 13 August 2021).

Recent automatic identification methods in taxonomic research using deep learning have mainly focused on modern organisms, with only a few considering fossils (mainly microfossils), including foraminifera (Zhong et al. 2017; Hsiang et al. 2019; Mitra et al. 2019; Carvalho et al. 2020; Marchant et al. 2020; Pires de Lima et al. 2020), radiolarians (Keçeli et al. 2017, 2018; Tetard et al. 2020), planktonic life forms (Al-Barazanchi et al. 2015, 2018; Li and Cui 2016), coccoliths (Beaufort and Dollfus 2004), diatoms (Urbankova et al. 2016; Bueno et al. 2017; Pedraza et al. 2017; Kloster et al. 2020; Lambert and Green 2020), pollen grains (Marcos et al. 2015; Kong et al. 2016; Sevillano and Aznarte 2018; Bourel et al. 2020; Romero et al. 2020), plants (Liu et al. 2018a; Kaya et al. 2019; Too et al. 2019; Ngugi et al. 2021), wild mammals (Villa et al. 2017; Norouzzadeh et al. 2018; Tabak et al. 2019), insects (Rodner et al. 2015; Martineau et al. 2017; Valan et al. 2019), and bones and teeth (Domínguez-Rodrigo and Baquedano 2018; Byeon et al. 2019; Hou et al. 2020; MacLeod and Kolska Horwitz 2020). However, various drawbacks remain to be addressed. First, most existing studies have focused on species-level automatic identification, but such a method can only be applied to a few common taxa in a specific clade. Hsiang et al. (2019) collected the largest planktonic foraminifera dataset, containing 34,000 images from 35 species (comprising most living planktonic foraminifera). They combined the efforts of more than 20 taxonomic experts to generate a rich and accurate dataset for deep learning. However, other fossil clades, such as benthic foraminifera, usually contain thousands of species, but recent studies only analyzed a limited number of taxa (Pires de Lima et al. 2020). Consequently, few experts can truly benefit from such studies. Second, although previous research provides publicly available codes and models, taxonomists may face usage difficulties owing to software limitations. Thus, providing an end-to-end framework is necessary and critical for the adoption of deep learning models. Third, experiments were conducted based on a dataset collected from personal collections (Mitra et al. 2019), research institution collections (Hsiang et al. 2019), public literature, and multiple sources (Liu and Song 2020; Pires de Lima et al. 2020). Although collecting data from various sources can partially compensate for data scarcity, tens of thousands or even millions of samples are usually needed to develop successful deep learning applications (Deng et al. 2009). Traditional data collection is unsuitable for constructing massive datasets unless geologists can standardize or digitize unstructured data scattered in paper records and personal hard disks (Wang et al. 2021).

To improve data collection, we used web crawlers to collect the largest Fossil Image Dataset (FID) currently available from the Internet, obtaining more than 415,000 images. We then leveraged the high performance of DCNNs to perform the automatic taxonomic identification of fossils. Rather than focusing on a particular fossil group, we first aimed to identify fossil clades. This approach can support research in the geosciences and help disseminate paleontological knowledge to the public. We also deployed the resulting model on a server for public access at www.ai-fossil.com.

Data and Methods

Data.—We collected the FID from the public Internet using web crawlers and then we manually checked the images and their corresponding labels. The uniform resource locators (URLs) of the images highlighted in this study are listed in Supplementary Table S1, and those of all the images in the FID are available at https://doi.org/10.5281/zenodo.6333970. Some collected images showed a large area of unnecessary backgrounds, which we manually cropped to preserve the fossiliferous regions only. Data on conodont, foraminifera, and trace fossils were supplemented from the literature, using the method demonstrated in Liu and Song (2020). We collected 415,339 images belonging to 50 clades in the FID (Fig. 2). The 50 clades consist of five superclades: (1) invertebrates: ammonoids, belemnites, bivalves, blastoids, brachiopods, bryozoans, chelicerates, corals, crinoids, crustaceans, echinoids, gastropods, graptolites, insects, myriapods, nautiloids, ophiuroids, sponges, starfish, stromatolites, and trilobites; (2) vertebrates: agnatha, amphibians, avialae, bone fragments, chondrichthyes, crocodylomorphs, mammals, mammalian teeth,

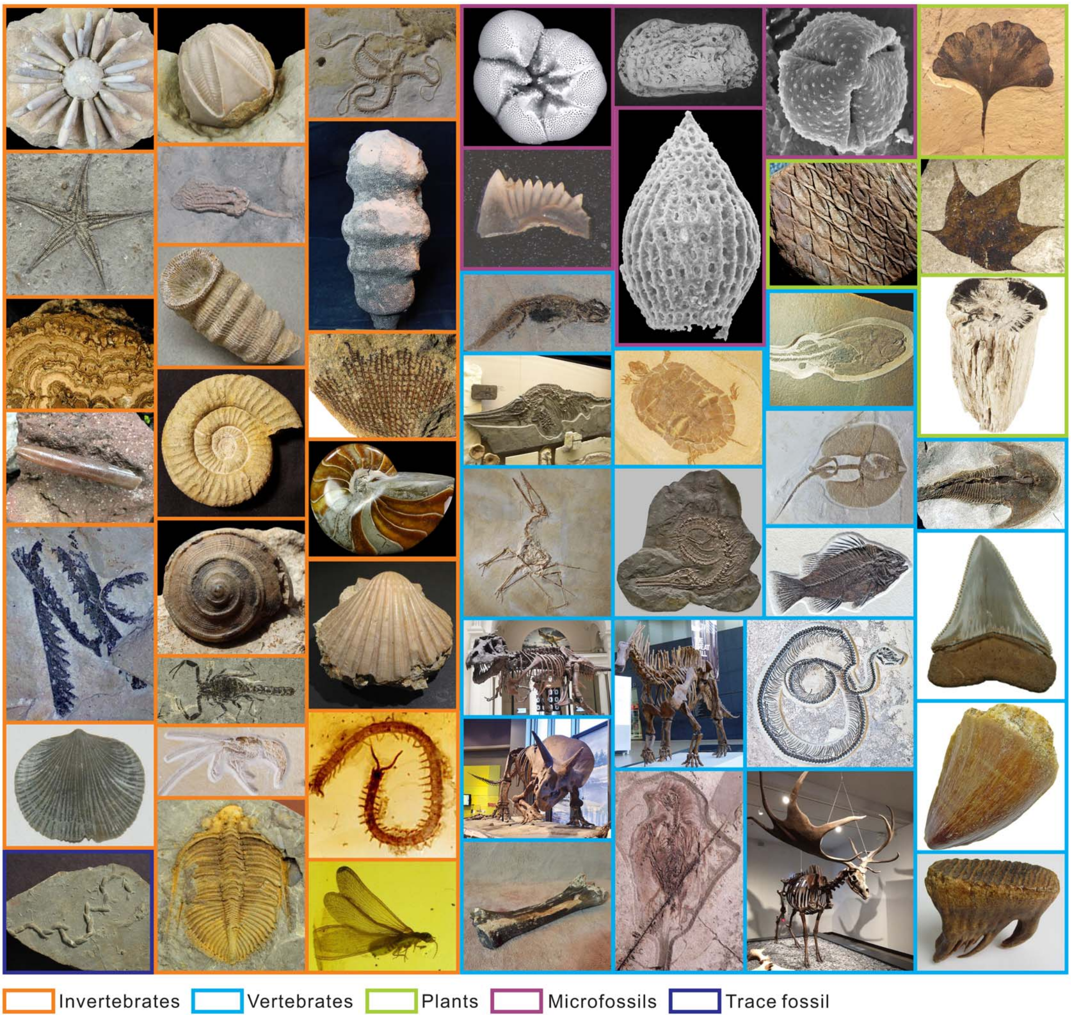

FIGURE 2. Example images of each class in our dataset, which contains 50 clades (Table 1). Specimens are not to scale. The source URLs of the images are provided in Supplementary Table S1.

marine reptiles, ornithischians, osteichthyes, placoderms, pterosaurs, reptilian teeth, sauropodomorphs, shark teeth, snakes, theropods, and turtles; (3) plants: angiosperms, gymnosperms, petrified wood, and pteridophytes; (4) microfossils: conodonts, foraminifera, ostracods, radiolarians, and spores or pollen; and (5) trace fossils.

Web Crawlers.—Web crawlers have been used to collect data from open web pages for data mining (Helfenstein and Tammela 2017; Lopez-Aparicio et al. 2018) and deep learning (Xiao et al. 2015). A web crawler is a programmed script or software that browses web pages systematically and automatically to retrieve specific information (Kausar et al. 2013) by sending requests for documents on servers to resemble a normal request. The script examines the returned data per web page to select useful information, such as image URLs in this study. We used search engines (e.g., Google and Bing) to collect fossil images by searching for keywords (e.g., "trilobite") and then downloaded the images and their associated URLs to a local storage site. We examined different keywords to download fossil images from different geological ages and regions. In addition, we removed duplicate images by applying algorithms such as AntiDupl.NET.

Computing Environments.—All analysis codes were executed on a Dell Precision 7920 Workstation running Microsoft Windows 10 Professional. The workstation was equipped with two Intel Xeon Silver 4216 processors, 128 GB of memory, and two NVIDIA GeForce GTX 2080Ti graphics processors (11 GB of memory per graphics processor). To implement the deep learning framework, we used TensorFlow v. 1.13.1 (Abadi et al. 2016) and Keras v. 2.2.4 (with TensorFlow backend; Chollet 2015) in Python v. 3.6.5. The required preinstalled Python libraries, algorithms for analysis, and model weights are available at https://github.com/XiaokangLiuCUG/Fossil_Image_Dataset. All the images used and their URLs are uploaded at https://doi.org/10.5281/zenodo.6333970.

Convolutional Neural Network.—A CNN is a supervised learning algorithm that requires images to be input with their corresponding labels for training. CNNs can handle image-based tasks, including image recognition (LeCun et al. 1998), object detection (Redmon et al. 2016), facial detection (Li et al. 2015), semantic segmentation (Long et al. 2015), and image retrieval (Babenko et al. 2014). Fukushima (1980) proposed a self-organized artificial neural network called the neocognitron, a predecessor of CNNs, that tolerates image shifting and deformation based on the work by Hubel and Wiesel (1962). LeCun et al. (1998) first used a backpropagation method in LeNet-5 to learn the convolution kernel coefficients directly from MNIST (Mixed National Institute of Standards and Technology database, which contains 60,000 grayscale images of handwritten digits) images. Subsequently, DCNNs were created, which usually contain dozens or even hundreds of hidden layers, such as VGG-16 (Simonyan and Zisserman 2014), GoogLeNet (Szegedy et al. 2015), ResNet (He et al. 2016), Inception-ResNet (Szegedy et al. 2017), and PNASNet (Liu et al. 2018b), and DCNNs are widely used in many domains.

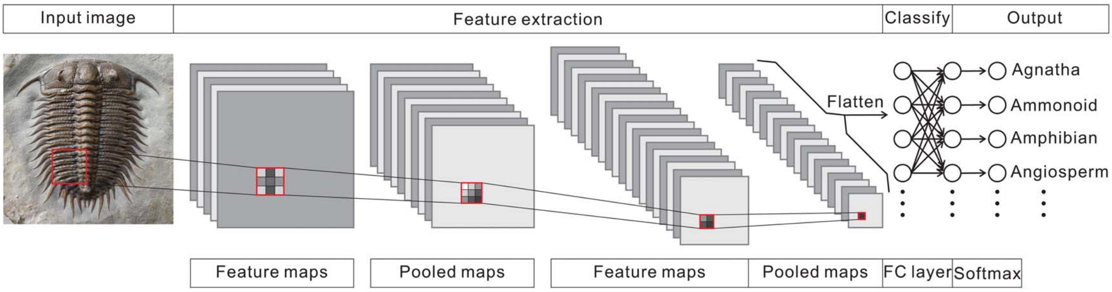

FIGURE 3. Schematic of a convolutional neural network, modified from Krizhevsky et al. (2012). FC layer, fully connected layer.

TABLE 1. Number of samples for the three subsets and each class.

Order	Clade	Training set (0.80)	Validation set (0.15)	Test set (0.05)	Total
0	Agnatha	1543	289	96	1928
1	Ammonoid	12,879	2414	804	16,097
2	Amphibian	2058	385	128	2571
3	Angiosperm	4060	761	253	5074
4	Avialae	8097	1518	506	10,121
5	Belemnite	3932	737	245	4914
6	Bivalve	6311	1183	394	7888
7	Blastoid	1734	324	108	2166
8	Bone fragment	14,982	2809	936	18,727
9	Brachiopod	5817	1090	363	7270
10	Bryozoan	2280	427	142	2849
11	Chelicerate	4561	855	285	5701
12	Chondrichthyes	2949	552	184	3685
13	Conodont	13,624	2554	851	17,029
14	Coral	12,298	2305	768	15,371
15	Crinoid	7213	1352	450	9015
16	Crocodylomorph	2743	514	171	3428
17	Crustacean	5458	1023	341	6822
18	Echinoid	8961	1679	559	11,199
19	Foraminifera	7262	1361	453	9076
20	Gastropod	6637	1244	414	8295
21	Graptolite	1849	346	115	2310
22	Gymnosperm	4721	884	294	5899
23	Insect	7725	1448	482	9655
24	Mammal	12,176	2282	760	15,218
25	Mammalian teeth	7082	1327	442	8851
26	Marine reptile	3514	658	219	4391
27	Myriapod	1238	232	77	1547
28	Nautiloid	3896	730	243	4869
29	Ophiuroid	2725	510	170	3405
30	Ornithischian	11,053	2072	690	13,815
31	Osteichthyes	11,267	2112	704	14,083
32	Ostracod	4507	844	281	5632
33	Petrified wood	13,798	2586	862	17,246
34	Placoderms	1624	304	101	2029
35	Pteridophyte	10,782	2021	673	13,476
36	Pterosaurs	3674	688	229	4591
37	Radiolarian	5174	970	323	6467
38	Reptilian teeth	10,771	2019	673	13,463
39	Sauropodomorph	4568	856	285	5709
40	Shark teeth	16,585	3109	1036	20,730
41	Snake	488	91	30	609
42	Sponge	2665	499	166	3330
43	Spore or pollen	7312	1370	456	9138
44	Starfish	2384	446	148	2978
45	Stromatolite	3559	667	222	4448
46	Theropod	16,504	3094	1031	20,629
47	Trace fossil	8048	1509	503	10,060
48	Trilobite	14,981	2808	936	18,725
49	Turtle	2249	421	140	2810
	Total	332,318	62,279	20,742	415,339

Inputs are the pixel matrix of an image, which is usually represented by a grayscale channel or red-green-blue channels for 2D images (e.g., trilobite image in Fig. 3), and its corresponding label. In general, a conventional CNN mainly consists of convolutional layers, pooling layers, a fully connected layer, and an output layer (Krizhevsky et al. 2012). An input image is successively convolved with learned filters in each layer, where each activation map can also be interpreted as a feature map. Then a nonlinear activation function performs a transformation to learn complex decision boundaries across images. The pooling layers are used for downsampling, considering that adjacent pixels contain similar information. Convolutional and pooling layers are usually combined and reused multiple times for feature extraction (Fig. 3). Then, the fully connected layer combines thousands of feature maps for the final classification. The output layer uses the *softmax* function in TensorFlow to generate a probability vector to represent the classification result. Cross-entropy is usually used as the objective function for measuring errors from predicted and true labels (Botev et al. 2013). The backpropagation of the gradient method was conducted to minimize the cross-entropy value and maximize the classification performance of the architecture (Rumelhart et al. 1986).

DCNNs contain massive parameters, and they should be trained on large datasets such as ImageNet (https://image-net.org; Deng et al. 2009), which contains more than 1.2 million labeled images from 1000 classes. Alternatively, transfer learning can be used for small training datasets (Tan et al. 2018; Brodzicki et al. 2020; Koeshidayatullah et al. 2020). In transfer learning, instead of training a CNN architecture from randomly initialized parameters, the parameters are obtained from pretraining on other recognition tasks with a large dataset for initialization. Thus, transfer learning can reduce computing costs, improve feature extraction, and accelerate the training convergence of the model. Transfer learning methods mainly include feature extraction and fine-tuning. For feature extraction, convolutional layers are frozen so that the parameters are not updated during training. In this study, we froze the shallow layers (half the network layers) to only train the deep layers. We also evaluated fine-tuning, in which pretrained parameters are used as initialization, and training was applied to all the network parameters with a small learning rate.

To increase the model's performance and the generalization ability of the evaluated DCNNs, we used data augmentation (Wang and Perez 2017), randomly adding noise to enhance the robustness of the algorithm against the contrast ratio, color space, and brightness, which are not the main identification characteristics of classes in fossil images (Shorten and Khoshgoftaar 2019). In addition, we applied random cropping, rotating, and resizing of the images using preprocessing packages in TensorFlow (Abadi et al. 2016) and Keras (Chollet 2015). Data augmentation allows expansion of the training set and partially mitigates overfitting (Wang and Perez 2017). In this study, we examined different combinations of data augmentation operations. We used the downscaled image for inference to reduce the image preprocessing time. The width and height of all the images were limited to 512 and then used to create TFRecord files.

Training.—We randomly split the collected FID into three subsets, as detailed in Table 1. The training set was used for determining the model parameters, while the validation set was used to adjust the hyperparameters (untrainable parameters) of the model and verify the performance during training, and the test set was used to evaluate the generalization ability of the final model. To test the influence of data volume on individual class accuracy, we also trained on a reduced FID, in which each class contained 1200 training images, and tested the final performance on the same test set. We used the top-1 accuracy (the true class matches the predicted label with the biggest possibility) and top-3 accuracy (the true class matches the predicted label for any of the three most probable classes) to measure the performance of the DCNN architectures. We evaluated three conventional DCNNs, namely Inception-v4 (Szegedy et al. 2017), Inception-ResNet-v2 (Szegedy et al. 2017), and PNASNet-5-large (Liu et al. 2018b), which have achieved excellent results on ImageNet. We ran 14 trials to optimize the performance of the models with different hyperparameter sets. In addition, we considered transfer learning and training from randomly initialized parameters. For transfer learning, we fine-tuned all trainable layers and froze the shallow layers (half of the DCNN layers) to only train the remaining layers, as detailed in Table 2. During training, the algorithm randomly fed a batch of images per iteration (step). An epoch was complete when all training images were fed to the architecture once, noting that dozens of epochs are usually required for the model convergence. The output logs can facilitate optimization. The test set was used to determine the ultimate

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

TABLE 2. Experiments for the three deep convolutional neural network (DCNN) architectures. For the "Load weights" column, pre-trained parameters were used for variable initialization (i.e., transfer learning). In the "Train layers" column, the settings of training/froze layers for Inception-v4 and Inception-ResNet-v2 follow the methods of Liu and Song (2020). For PNASNet-5-large, the trainable layers include cell_6, cell_7, cell_8, cell_9, cell_10, cell_11, aux_7, and final_layer in Liu et al. (2018b). "DA with RC" shows data augmentation with the random crop, the random cropped image covers 0.4–1 range of the original image (except experiment 7, which used a range of 0.65–1). Other data augmentation methods follow the methods of Szegedy et al. (2017) and Liu et al. (2018b). All experiments used batch normalization, dropout (with 0.8), and Adam optimizer. The input size of Inception-v4 and Inception-ResNet-v2 is 299 × 299, and that of PNASNet-5-large is 331 × 331. The decay rate for experiment 6 is 0.96 when training epochs <15 and 0.9 when training epochs ≥15. Experiment 13 was trained on the reduced Fossil Image Dataset (FID). During the training processing, we printed the output (including train/validation loss and accuracy) for each 1000 iterations and tried the model's performance for each two epochs. The maximum training/validation accuracy and minimum training/validation loss have the best results among all outputs. Similarly, the maximum top-1/top-3 test accuracies have the best performance of the whole training process.

Order	Architecture	Batch size	Load weights	Froze layers	Train layers	Start learning rate	Decay step	Decay rate	DA with RC	Epochs run	Max. train accuracy	Min. training loss	Max. validation accuracy	Min. validation loss	Max. top-1 test accuracy	Max. top-3 test accuracy
1	Inception-v4	55	Yes	Yes	Half layers	0.00001	3000	0.96	No	24	100.00%	0.3666	100.00%	0.3810	88.71%	96.18%
2	Inception-v4	64	Yes	Yes	Half layers	0.0001	epoch/2	0.96	Yes	40	100.00%	0.3637	100.00%	0.4055	88.96%	96.15%
3	Inception-v4	64	Yes	Yes	Half layers	0.0001	epoch/2	0.96	No	40	100.00%	0.3095	98.44%	0.3598	88.87%	96.03%
4	Inception-v4	64	Yes	Yes	Half layers	0.0001	epoch	0.95	Yes	40	100.00%	0.3661	96.88%	0.4538	88.47%	96.00%
5	Inception-v4	64	Yes	Yes	Half layers	0.0001	epoch	0.95	No	40	100.00%	0.3036	96.88%	0.5181	88.15%	95.83%
6	Inception-v4	64	Yes	Yes	Half layers	0.00001	epoch	0.9(6)	Yes	40	100.00%	0.4111	95.31%	0.5264	87.99%	95.90%
7	Inception-v4	64	Yes	Yes	Half layers	0.00001	epoch*2	0.96	Yes	28	100.00%	0.3775	95.31%	0.5421	87.46%	95.86%
8	Inception-v4	64	Yes	Yes	Half layers	0.00001	epoch*3	0.96	No	40	100.00%	0.3572	96.88%	0.5964	86.91%	95.17%
9	Inception-v4	32	No	No	All layers	0.0005	epoch/2	0.96	Yes	40	100.00%	0.3653	96.88%	0.3659	82.85%	93.04%
10	Inception-ResNet-v2	64	Yes	Yes	Half layers	0.0001	epoch/2	0.96	Yes	60	100.00%	0.3926	96.88%	0.5083	90.12%	96.50%
11	Inception-ResNet-v2	32	No	No	All layers	0.0005	epoch/2	0.98	Yes	40	100.00%	0.3529	100.00%	0.4117	79.10%	91.32%
12	Inception-ResNet-v2	32	Yes	Yes	All layers	0.001	epoch/2	0.98	Yes	60	100.00%	0.2736	96.88%	0.3947	83.33%	93.70%
13	Inception-ResNet-v2	64	Yes	Yes	Half layers	0.0001	epoch/2	0.96	Yes	60	100.00%	0.5054	90.62%	0.9189	82.97%	92.57%
14	PNASNet-5-large	32	Yes	Yes	Half layers	0.0004	epoch/2	0.96	Yes	40	100.00%	0.5164	100.00%	0.5602	88.03%	95.80%

performance of each DCNN. The main experiments were performed using Inception-v4 to optimize the hyperparameters, and then we trained on the other two architectures. We did not try all possible experiments to optimize the hyperparameters, because we evaluated the identification of biotic and abiotic grains in thin sections in the experiment by Liu and Song (2020). Instead, we used the main settings and tried to optimize several critical hyperparameters, such as those for transfer learning, fine-tuning, learning rate, and data augmentation operations.

Evaluation Metrics.—Several metrics were calculated to evaluate the performance of each experiment on the test set. Among them, recall measures the ratio of correctly predicted positive labels against all observations in the actual class, that is, true position/(true positive + false negative); precision measures the ratio of correctly predicted positive labels to the total predicted positive observations, that is, true positive/(true positive + false positive); and the F_1 score is a comprehensive index that is the harmonic mean, which is calculated as 2 × precision × recall/(precision + recall) (Fawcett 2006). The receiver operating characteristic (ROC) curve measures the sensitivity of the models to the relative distribution of positive and negative samples within a class based on the analysis of the output probabilities of all samples. The area under the ROC curve (AUC) represents the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative example (Fawcett 2006). Macro-averaged AUC calculates metrics for all classes and finds their unweighted mean. This metric does not take label imbalance into account. Micro-averaged AUC calculates metrics globally by considering each element of the label indicator matrix as a label, which is a weighted value based on the relative frequencies of each class (Sokolova and Lapalme 2009).

Visualization of Feature Maps.—Although CNN architectures have demonstrated high efficiency in solving complex vision-based tasks, they are regarded as black boxes that hinder explanation of their internal workings. Accordingly, methods to explain the workings of CNNs have been developed (Selvaraju et al. 2017; Fukui et al. 2019). In this study, we aim to visualize the characteristics that DCNNs learn to perform identification on images from the collected FID. Selvaraju et al. (2017) proposed a method called gradient-weighted class activation mapping (Grad-CAM) to visualize class discrimination and locate image regions that are relevant for classification. Feature visualization uses the output of the final convolutional layer (spatially pooled by global average pooling) because that layer contains spatial information (high-level visual constructs) lost in the last fully connected layer (Fig. 3). Accordingly, we used a 3D (1536 × 8 × 8) matrix obtained from the Inception-ResNet-v2 architecture (see the schematic of Inception-ResNet-v2 in Supplementary Fig. S1). Grad-CAM uses the gradients of any target label flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the maximum probability label. Nevertheless, Grad-CAM cannot highlight fine-grained details such as pixel-space gradient visualization methods (Selvaraju et al. 2017). To overcome this limitation, we also applied guided Grad-CAM by pointwise multiplication of the heat map using guided backpropagation to obtain the high-resolution and concept-specific images of the most representative features (Selvaraju et al. 2017). Guided Grad-CAM may only capture the most discriminative part of an object (pixels or regions with large gradients), and its threshold may not highlight a complete object, unlike saliency maps (Simonyan et al. 2013). Grad-CAM and guided Grad-CAM are based on gradient backpropagation (Selvaraju et al. 2017). In this study, we utilized Grad-CAM, guided Grad-CAM, and the extracted feature maps to perform visual explanation. These visualization maps show which areas and features are important for DCNN architectures to identify different fossils.

To unveil interactions in different specimens and groups, we applied t-distributed stochastic neighbor embedding (t-SNE) to visualize the feature maps extracted by the Inception-ResNet-v2 architecture. This type of embedding allows visualizing high-dimensional data through the dimensionality reduction of

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

each data point to two or three dimensions, which was presented by Maaten and Hinton (2008). It starts by converting high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. Then, the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and high-dimensional data is minimized (Maaten and Hinton 2008). We randomly selected 40 specimens for each clade and visualized the global average pooling layer (a vector with the shape of 1536) from layer conv_7b (3D matrix with a dimension of 1536 × 8 × 8). We also applied another dimensionality reduction method called uniform manifold approximation and projection (UMAP) (McInnes et al. 2018) to verify our results. UMAP is also used for nonlinear dimension reduction and preserves more of the global structure with superior run-time performance. Furthermore, UMAP has no computational restrictions on embedding dimensions.

Results

Three DCNN architectures were performed similarly on FID, but the hyperparameter settings influenced the performance. Among them, Inception-v4 achieved 0.89 top-1 accuracy and 0.96 top-3 accuracy on the test set, corresponding to a minimum training loss of 0.36 and a minimum validation loss of 0.41 (analysis 2 in Table 2). The highest top-1 and top-3 accuracies are 0.90 and 0.97, respectively, on the test set obtained by the Inception-ResNet-v2 architecture (analysis 10 in Table 2). For PNASNet-5-large architecture, we conducted one fine-tuning experiment, obtaining 0.88 top-1 accuracy and 0.96 top-3 test accuracy, representing a slightly inferior performance compared with the other two architectures. The minimum training/validation measures the behavior of the model on training and validation sets during the training process. They are also affected by other hyperparameters, such as batch size and the number of training layers. In Table 2, training all layers with a batch size of 32 is more likely to result in minimal validation loss (analyses 9, 11, and 12). Overall, transfer learning outperforms random parameter initialization, and fine-tuning of deep layers of the DCNN is more effective by approximately 7% than fine-tuning the entire DCNN. The transfer learning method accelerates model convergence and provides a stable loss during training (Fig. 4). In addition, a frequent learning rate decay improves the identification performance. Among the data augmentation operations, applying random cropping to the training images promotes the learning of local characteristics of fossils and improves the generalization performance (Table 2).

The results of analysis 10 (Table 2) indicate an overall accuracy of 0.90 for the validation and test images, corresponding to an unweighted average recall of 0.88 ± 0.08, unweighted average precision of 0.89 ± 0.07, and unweighted average F_1 score of 0.88 ± 0.08. For recall, the images of conodonts, radiolarians, shark teeth, trilobites, and spores or pollen achieved the highest values of 0.99, 0.97, 0.96, 0.96, and 0.96, respectively, whereas bone fragments (0.80), trace fossils (0.78), agnathans (0.77), bryozoans (0.63), and sponges (0.52) showed the lowest recall. The precision and F_1 score showed similar trends for recall within the classes (Table 3). The identification achieved a higher accuracy for microfossils than for the other clades. In addition, we trained the Inception-ResNet-v2 on a reduced FID and used the same hyperparameters in analysis 10, demonstrating a top-1 accuracy of 0.83 and an unweighted average recall of 0.83 ± 0.09. For the individual class accuracy, most of the clades benefited from the larger training images, such as bone fragments (0.20 higher compared with that of reduced FID; Fig. 5), corals (0.15), pteridophytes (0.15), and trace fossils (0.16), while fossil clades with fewer training images in the FID exhibit higher accuracy in reduced FID, such as bryozoans (0.04 higher compared with the FID), myriapods (0.06), and sponges (0.06; Supplementary Table S2). We used random oversampling to expand the training images of the snake fossils to 1200, which improved the performance by 0.03 compared with the FID, whereas its performance was reduced after training from 40 epochs to 60 epochs in the reduced FID.

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

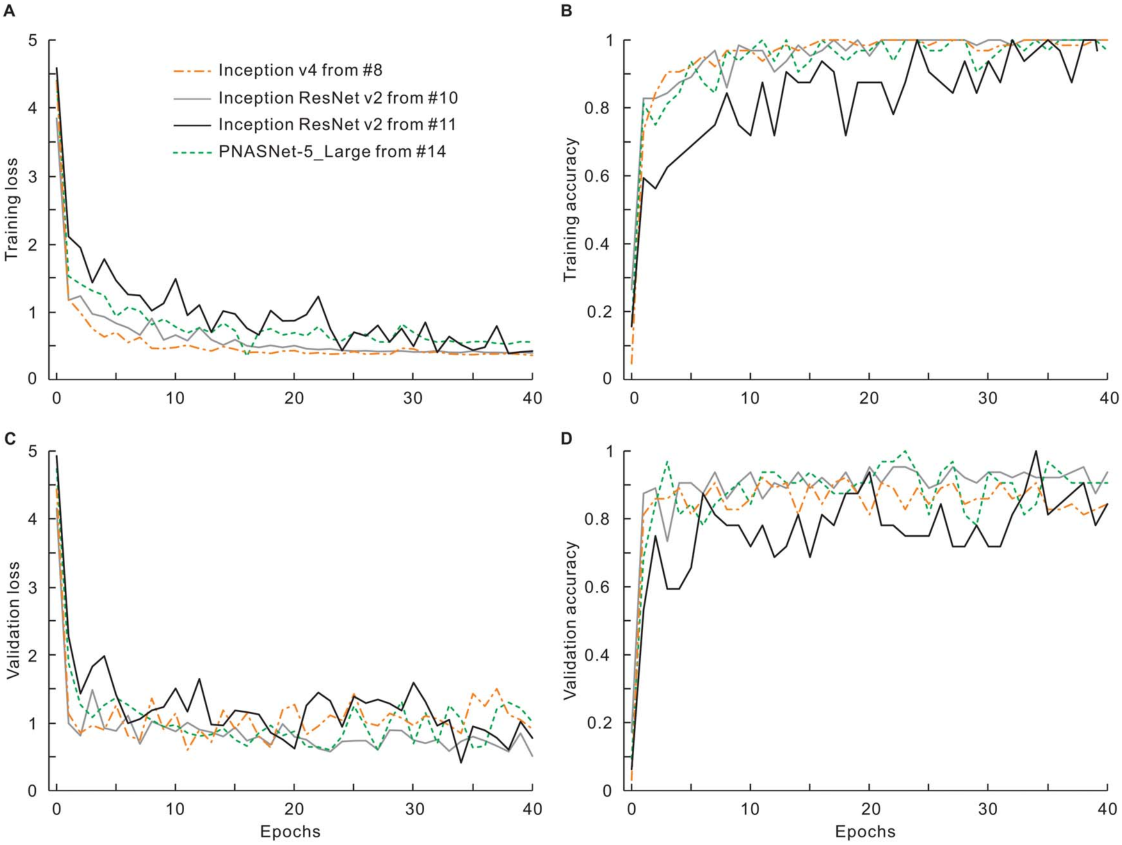

Figure 4. Curves demonstrate the (A) training loss, (B) training accuracy, (C) validation loss, and (D) validation accuracy of three deep convolutional neural network architectures during the training process. Experiments 8, 10, 11, and 14 are from Table 2. The fluctuations of the validation loss/accuracy may result in a higher learning rate. With a lower learning rate, more training epochs could smooth the curves and improve the accuracy, but it would also take longer to train the model, considering it currently takes 40–100 hours to train 40 epochs (depending on whether deep half layers or all layers were fine-tuned).

Given the false positives in the main categories of microfossils, plants, invertebrates, and vertebrates, the DCNN tends to learn the main morphological features of fossils. Specifically, the misidentified images usually occurred in a class with a higher morphological disparity, or it was difficult to learn the unique characteristics of a class due to the image quality, data volume, and some other adverse aspects, especially for sponges, bryozoans, bone fragments, and trace fossils. Sponges were frequently misidentified as corals (rate of 0.14), bone fragments (0.06), and trace fossils (0.04). Bryozoans were mostly misidentified as corals (0.12), trace fossils (0.05), and sponges (0.05). Mammalian tooth specimens were misidentified as bone fragments (0.08), reptilian teeth (0.04), and mammals (0.02) (Table 3). The confusion matrix of 50 clades is provided in Supplementary Table S2. Several groups with low identification performance were often confused with one another, such as sponges, bryozoans, trace fossils, and corals. In addition, fossil fragments can undermine the identification of other clades, such as bone fragments, teeth, and petrified wood.

Figure 6 demonstrates the average values of 50 clades and the five highest- and lowest-performing clades based on the ROC curve from the validation and test sets. The AUC result shows that although some of the samples were predicted incorrectly, they have a much higher sensitivity than negative samples, such as samples from sponges and bryozoans. The positive rates rapidly increased when the false-

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

TABLE 3. Optimum performance from experiment 10 of Inception-ResNet-v2, which analyzed validation and test datasets to reduce occasional fluctuations in data.

						Top-1		Top-2		Top-3	
Order	Clade	Total	Precision	F_1 score	Recall	error	Label	error	Label	error	Label
13	Conodont	3406	99.32%	99.22%	99.12%	0.18%	Bone fragment	0.15%	Radiolarian	0.12%	Foraminifer
37	Radiolarian	1294	97.00%	97.15%	97.30%	0.54%	Foraminifer	0.54%	Sponge	0.46%	Spore or pollen
40	Shark teeth	4144	96.48%	96.48%	96.48%	1.04%	Reptilian teeth	0.89%	Bone fragment	0.34%	Mammalian teeth, Chondrichthyes
48	Trilobite	3743	95.17%	95.79%	96.42%	0.48%	Stromatolite	0.40%	Trace fossil	0.27%	Crinoid
43	Spore or pollen	1827	98.32%	97.32%	96.33%	1.48%	Foraminifer	0.60%	Radiolarian	0.22%	Ostracod
18	Echinoid	2238	95.21%	95.55%	95.89%	0.54%	Mammal	0.54%	Mammal	0.40%	Gastropod
1	Ammonoid	3219	94.64%	95.08%	95.53%	1.06%	Gastropod	0.50%	Bivalve	0.43%	Nautiloid
31	Osteichthyes	2810	94.04%	94.77%	95.52%	0.68%	Chondrichthyes	0.46%	Bone fragment	0.43%	Crustacean
32	Ostracod	1125	97.72%	96.40%	95.11%	0.98%	Foraminifer	0.62%	Spore or pollen	0.44%	Trilobite, Conodont, Bivalve
19	Foraminifer	1816	94.03%	94.26%	94.49%	0.61%	Radiolarian	0.50%	Trace fossil	0.44%	Sponge, Gastropod
24	Mammal	3045	91.35%	92.51%	93.69%	2.17%	Bone fragment	0.76%	Avialae	0.72%	Ornithischian
30	Ornithischian	2757	92.55%	92.90%	93.25%	3.08%	Theropod	0.83%	Sauropodomorph	0.62%	Mammal
23	Insect	1930	94.39%	93.80%	93.21%	2.44%	Chelicerate	0.88%	Myriapod	0.62%	Angiosperm
39	Sauropodomorph	1140	93.39%	93.14%	92.89%	3.95%	Theropod	2.02%	Ornithischian	0.61%	Mammal
5	Belemnite	978	94.78%	93.75%	92.74%	1.23%	Reptilian teeth	1.02%	Bone fragment	0.92%	Nautiloid
33	Petrified wood	3438	91.62%	92.10%	92.58%	1.34%	Bone fragment	0.93%	Coral	0.81%	Gymnosperm, Stromatolite
46	Theropod	4121	90.71%	91.58%	92.45%	2.86%	Avialae	2.01%	Ornithischian	0.66%	Sauropodomorph
29	Ophiuroid	681	90.49%	91.35%	92.22%	2.79%	Starfish	1.62%	Crinoid	1.03%	Trace fossil
17	Crustacean	1363	92.87%	92.28%	91.71%	1.17%	Chelicerate	0.88%	Osteichthyes	0.59%	Coral
36	Pterosaurs	917	90.69%	91.04%	91.38%	2.73%	Avialae	1.64%	Theropod	1.09%	Bone fragment
38	Reptilian teeth	2690	90.49%	90.54%	90.59%	3.87%	Bone fragment	1.86%	Mammalian teeth	1.26%	Shark teeth
11	Chelicerate	1141	89.52%	89.68%	89.83%	4.73%	Insect	0.70%	Trilobite	0.61%	Crustacean
14	Coral	3071	87.20%	88.28%	89.38%	1.89%	Sponge	1.24%	Bryozoan	1.07%	Bone fragment
35	Pteridophyte	2693	88.48%	88.74%	89.01%	2.90%	Gymnosperm	1.49%	Trace fossil	0.89%	Angiosperm
28	Nautiloid	973	93.22%	91.06%	89.00%	4.42%	Ammonoid	1.54%	Crinoid	1.03%	Belemnite
4	Avialae	2022	86.66%	87.48%	88.33%	6.33%	Theropod	1.29%	Pterosaurs	1.19%	Mammal
20	Gastropod	1659	87.62%	87.75%	87.88%	2.71%	Bivalve	1.21%	Ammonoid	0.96%	Coral
27	Myriapod	310	90.37%	89.03%	87.74%	2.90%	Insect	1.94%	Chelicerate	0.97%	Pteridophyte, Crinoid
15	Crinoid	1802	85.00%	85.72%	86.46%	1.94%	Coral	1.50%	Trace fossil	0.83%	Bone fragment, Sponge
2	Amphibian	514	87.38%	86.78%	86.19%	3.31%	Osteichthyes	1.95%	Bone fragment	1.17%	Avialae
12	Chondrichthyes	738	88.19%	87.11%	86.04%	3.25%	Osteichthyes	2.85%	Bone fragment	0.81%	Shark teeth
9	Brachiopod	1454	84.75%	85.35%	85.97%	6.12%	Bivalve	1.03%	Trilobite	0.76%	Crinoid
49	Turtle	563	89.96%	87.92%	85.97%	2.84%	Bone fragment	1.60%	Trace fossil	0.89%	Mammal, Ornithischian, Marine reptile
41	Snake	122	87.39%	86.31%	85.25%	4.10%	Mammal	1.64%	Trace fossil	1.64%	Ornithischian, Marine reptile
34	Placoderms	403	85.54%	85.32%	85.11%	2.48%	Chelicerate	2.48%	Chelicerate	1.24%	Osteichthyes
44	Starfish	594	89.70%	87.29%	85.02%	5.05%	Ophiuroid	2.86%	Trace fossil	1.52%	Crinoid
7	Blastoid	431	87.95%	86.29%	84.69%	3.25%	Crinoid	2.55%	Brachiopod	1.39%	Foraminifer
26	Marine reptile	878	87.12%	85.50%	83.94%	2.28%	Theropod	2.05%	Crocodylomorph	1.94%	Bone fragment
6	Bivalve	1573	81.21%	81.95%	82.71%	6.55%	Brachiopod	1.84%	Gastropod	1.27%	Trace fossil

Table 3. Continued.

Order	Clade	Total	Precision	F_1 score	Recall	Top-1 error	Label	Top-2 error	Label	Top-3 error	Label
22	Gymnosperm	1179	82.68%	82.65%	82.61%	8.23%	Pteridophyte	3.39%	Angiosperm	1.78%	Petrified wood
21	Graptolite	463	83.33%	82.70%	82.07%	3.46%	Trace fossil	3.02%	Pteridophyte	1.30%	Bryozoan, Angiosperm
16	Crocodylomorph	685	87.54%	84.70%	82.04%	4.23%	Theropod	4.09%	Marine reptile	2.34%	Bone fragment
45	Stromatolite	889	83.24%	82.39%	81.55%	5.29%	Petrified wood	3.49%	Trilobite	1.80%	Trace fossil
3	Angiosperm	1013	85.17%	83.06%	81.05%	4.05%	Gymnosperm	3.26%	Pteridophyte	2.17%	Osteichthyes
25	Mammalian teeth	1770	82.58%	81.59%	80.62%	8.36%	Bone fragment	4.18%	Reptilian teeth	1.64%	Mammal
8	Bone fragment	3743	77.97%	78.98%	80.02%	3.37%	Mammalian teeth	2.94%	Mammal	1.52%	Reptilian teeth
47	Trace fossil	2012	79.07%	78.60%	78.13%	2.14%	Bone fragment	1.74%	Pteridophyte	1.59%	Sponge
0	Agnatha	386	85.47%	81.41%	77.72%	3.37%	Placoderms	2.85%	Trace fossil	2.59%	Bone fragment
10	Bryozoan	570	71.54%	66.79%	62.63%	12.46%	Coral	5.26%	Trace fossil	4.74%	Sponge
42	Sponge	667	59.62%	55.31%	51.57%	13.64%	Coral	6.15%	Bone fragment	4.50%	Trace fossil

positive rates were still lower. The micro average ROC and macro average ROC were similar considering the volume of the validation and test sets and the moderate data imbalance. The ROC curves of the 50 clades considered in this study are shown in Supplementary Figure S2.

Discussion

Performance Analysis.—We used the models that were first trained on the ImageNet (Deng et al. 2009) for transfer learning on the FID, and they exhibited outstanding results, which indicates that pretraining has been effective for applications in different recognition tasks (Wang et al. 2017; Willi et al. 2019), despite the fossil images being considerably different from the ImageNet samples (Yosinski et al. 2014). Hence, feature extraction using a CNN has a high generalization ability in different recognition tasks (Zeiler and Fergus 2014; Pires de Lima et al. 2020). Our approach, which froze half of the network layers as feature extractors and trained the remaining layers, provides the best performance. Transfer learning is also susceptible to overfitting, which may lead to experiments in which fine-tuning of all layers is inferior to fine-tuning of the deeper half-layers. We explored data augmentation, dropout, regularization, and early stopping methods to prevent such a situation. The first two methods are effective. We found that a frequent learning rate decay and large training batch contribute to faster convergence and high accuracy. The optimization of hyperparameters is usually empirical (Hinz et al. 2018) and becomes less effective as training proceeds. Compared with the model trained on reduced FID, most of the individual class accuracies linearly increased with the volume of training images, and the correlation coefficient was 0.73. Some of the clades with fewer than 3000 training images in the FID exhibited inferior performance on the reduced FID, but all classes with more than 3000 training images improved their accuracy on the complete FID. Imbalanced data caused the algorithm to pay more attention to categories with more training data. Sampling methods are typically used for imbalanced learning (He and Garcia 2009). In reduced FID, we used random undersampling

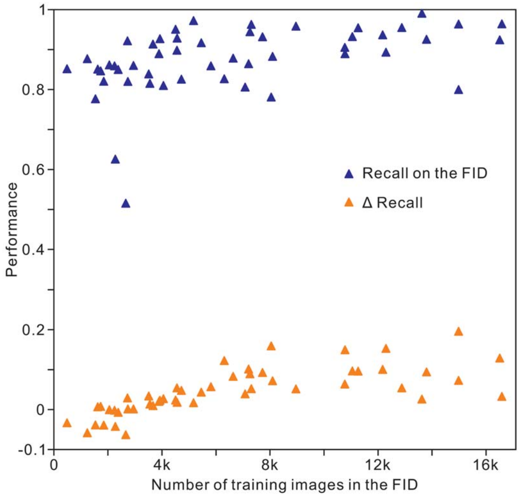

FIGURE 5. Distribution of individual clade recall with the volume of training images in the Fossil Image Dataset (FID). Δrecall equals accuracy on the FID minus accuracy on the reduced FID.

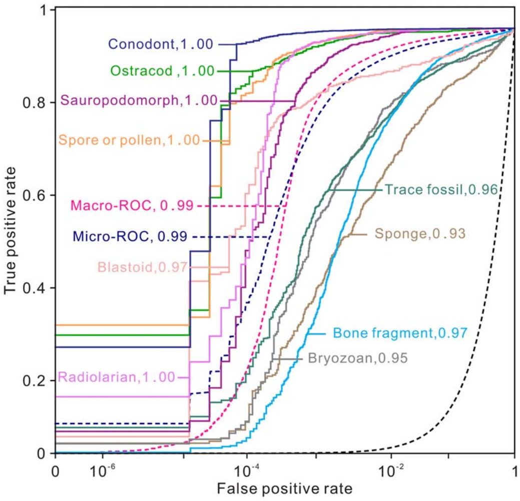

FIGURE 6. Receiver operating characteristic (ROC) curves of an average of 50 clades (dashed curves), the five highest, and five lowest classes from the validation and test datasets. AUC describes the area under the ROC curve. Ideally, an area close to 1 is the best scenario. Black dashed line comprises 0.5 ROC space, indicating a random prediction.

to remove the majority clades and used random oversampling to expand the minority clades (oversampling method only used for snake fossils). The results show that removing images from the majority leads to missing important content from the majority of clades, whereas oversampling is effective to a certain extent. Moreover, the dataset quality is important for accurate identification. The microfossils performed with high identification accuracy, because most of the specimens were collected from publication plates that provide images with less background and noise. Some fossils with poor preservation and few samples available performed poorly, such as bryozoans and sponges. The large intraclass morphological diversity of a clade (e.g., trace fossils and bone fragments) also undermined the identification accuracy, because it is difficult for the DCNN architecture to extract discriminative characteristics. For instance, trace fossils comprise coprolites, marine trace fossils, terrestrial footprints, and reptilian egg fossils, thus involving fickle morphologies and characteristics. Considering the data imbalance between these classes and with other groups, we did not further subdivide trace fossils into the four abovementioned classes.

Visual Explanation of Fossil Clade Identification.—Although irrelevant background noise may pollute test images, the DCNN architecture can extract representative areas. Nevertheless, the most discriminating areas are generally local features in fossil images, as shown in the Grad-CAM results in Figure 7. The red (blue) regions correspond to a high (low) score for predicting the label in Grad-CAM considering the average activation of the 1536 feature maps from Inception-ResNet-v2. Normally, the attention area of each feature map can be focused on the limited or unique characteristics (Selvaraju et al. 2016). A similar pattern is observed in the feature maps (Fig. 8), and some of the feature maps highlight the umbilicus, ribs, and inner whorl of the ammonoid. In particular, Inception-ResNet-v2 identifies different structures for each feature map in deep layers. (e.g., layers of mixed_7a and mixed _7b). Some feature maps are highly activated in a region limited to several pixels, indicating feature maps focus on specific spatial positions of the original image containing representative high-level structures. This phenomenon can be explained by the inherent characteristic of DCNNs, which compress the size (length and width) and increase the number (channels) of feature images through

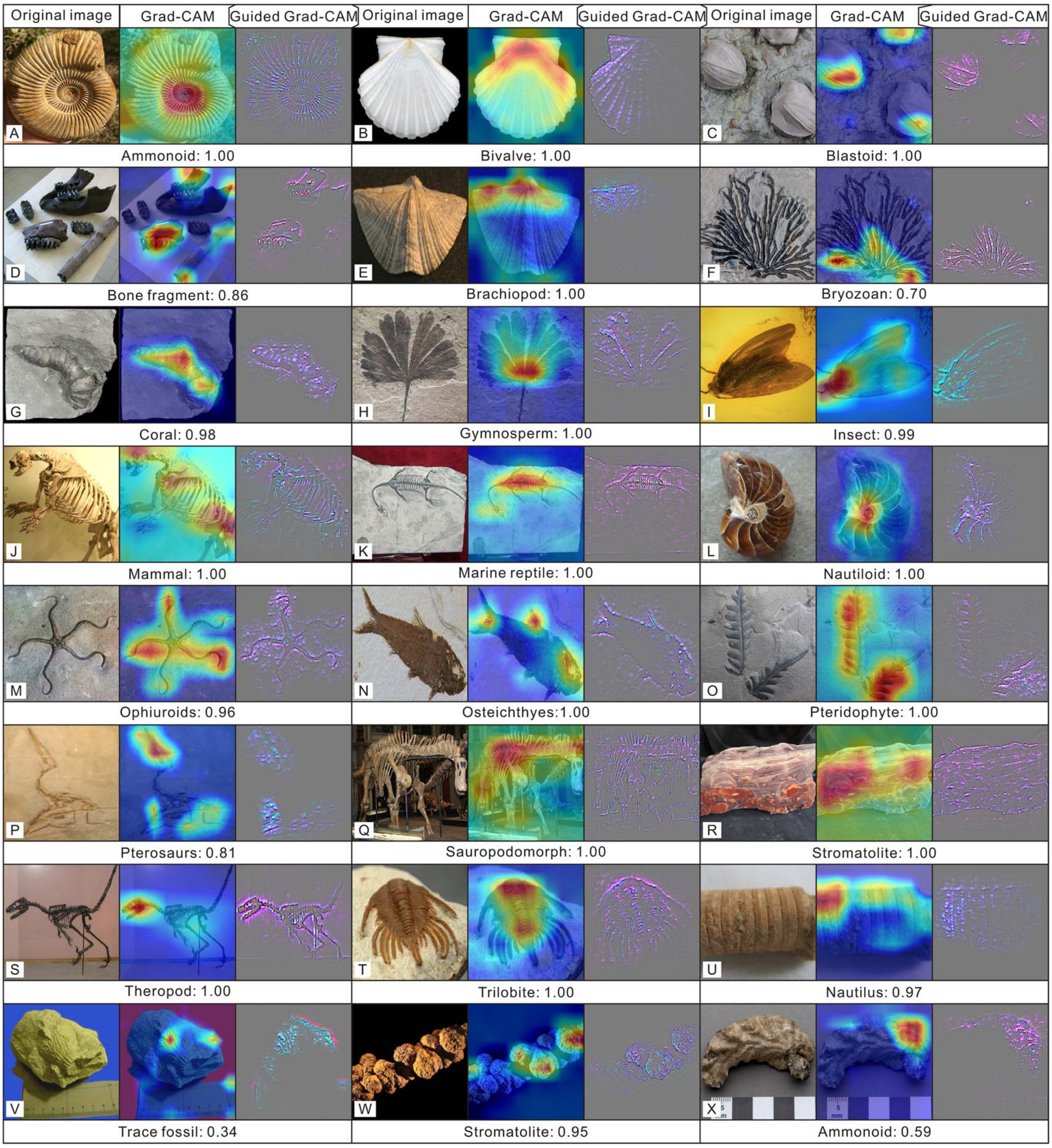

FIGURE 7. Visual explanation of samples from the test set, including the original image, gradient-weighted class activation mapping (Grad-CAM) fused with the original image, and guided Grad-CAM. The lower rectangle shows the predicted label and its probability. U–X were predicted incorrectly, and their true labels are crinoid, sponge, trace fossil, and bivalve, respectively. The red (blue) regions correspond to a high (low) score for predicting contribution in Grad-CAM. Specimens are not to scale. The image URLs are provided in Supplementary Table S1.

repeated convolutions. The feature maps from different convolutional layers demonstrate that shallow layers are sensitive to low-level features, such as brightness, edges, curves, and other conjunctions (layers conv2d_3 and conv2d_5 in Fig. 8) (Zeiler et al. 2011; Zeiler and Fergus 2014). Conversely, deeper layers detect complex invariant characteristics within classes or capture similar textures by combining some low-level features. The activation of a single feature map focused on a small area that should be class discriminative (Zeiler and Fergus 2014; Selvaraju et al. 2017), such as figures H, I, and J from layer mixed_6a in Figure 8.

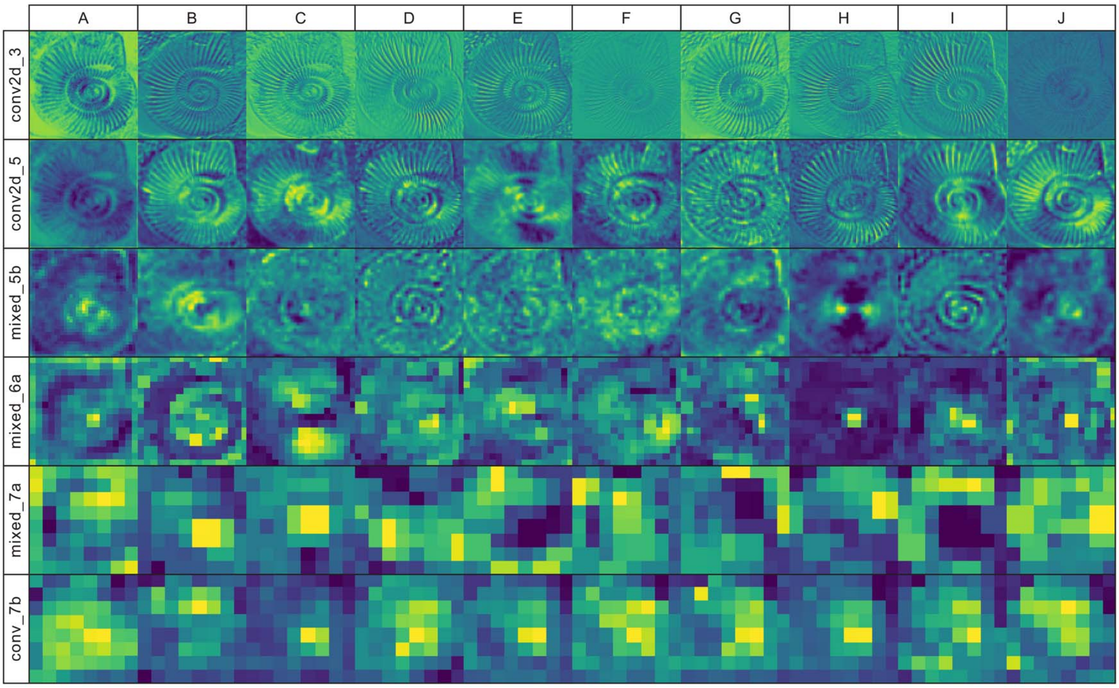

FIGURE 8. Visualization of the feature maps of different layers from Inception-ResNet-v2. From convd2_3 to mixed _7b (Supplementary Fig. S1), layers become deeper. First column of each layer is the averaged feature map (A), and the remaining column feature maps are nine examples of this layer (B–J). The dimensions of convd2_3, convd2_5, mixed_5b, mixed_6a, mixed_7a, and conv_7b are 147 × 147 × 64, 71 × 71 × 192, 35 × 35 × 320, 17 × 17 × 1088, 8 × 8 × 2080, and 8 × 8 × 1536, respectively. A schematic of the Inception-ResNet-v2 architecture is provided in Supplementary Fig. S1. Yellow (blue) pixels correspond to higher (lower) activations.

The activation patterns and feature maps show that the DCNN architecture can capture the fine-grained details of different fossils. For instance, the discriminative features extracted from ammonoids are mainly concentrated in the circular features merging from the umbilicus, especially for shells decorated with strong ribs (Fig. 7A). The spiral nautiloids highlight the area that contains similar characteristics. The feature maps demonstrate that the umbilical area is highlighted even in the middle layers (e.g., layers of mixed_5b and mixed_6a). Gastropod spires and apices are usually highlighted in Grad-CAM, considering that they contain curves and structures with large curvatures. For bivalves and brachiopods, the ornamented features of shells, such as growth lines and radial ribs, were extracted as representative features (Fig. 7B,E). Surprisingly, the DCNN seems to pay more attention to the dorsal or beak area of the shells, especially for brachiopods. These areas not only reflect significant differences between bivalves and brachiopods but also attract the attention of taxonomists. Therefore, DCNNs can capture the characteristics of fossils that are of interest to paleontologists. For arthropods, DCNN can highlight two pincers in crayfish, whereas it mainly concentrates on the body and tail of prawns and the head and wings of insects. The DCNN architecture provides outstanding performance in identifying vertebrate fossils. Grad-CAM emphasizes the skull and trunk of the vertebrate fossils, where distinguishing characteristics are located, especially for terrestrial vertebrates. This result may partially be attributed to some images in the FID showing only skulls. Class discrimination may result from a combination of several localized features. For instance, images of osteichthyes frequently show high activation on the skull, fins, and caudal fin areas (Fig. 7N). This phenomenon is also determined by Inception-v3 trained on ImageNet. Olah et al. (2018) built

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

blocks of interpretability for an image that contains a Labrador retriever and tiger cat. The corresponding attribution maps reflected semantic features in deep layers, such as the dog's floppy ears, snout, and face and the cat's head. Similarly, the implemented DCNN can extract multiple targets from an image. Three areas are highlighted, corresponding to three blastoid specimens in Figure 7, despite some of the fossils being fragments. This result was confirmed in the images from bone fragments, ophiuroids, and pteridophytes. For the misrecognized images, we cannot fully interpret classification based on the visualization maps, but Grad-CAM and guided Grad-CAM provide activation areas with a higher morphological similarity in the identified error labels for images such as those of a crinoid stem and straight-shelled nautiloid. The clump structure of coprolite (trace fossil) is similar to that of modern globular stromatolites in Shark Bay. In summary, a DCNN can effectively extract features from images. Some complicated texture features (e.g., complex curves and boundaries) and clade-specific structures are paid more attention to and used for decision making. The class activation and feature maps can help humans partially understand how they work on fossil images.

Selvaraju et al. (2017) demonstrated four visualization maps, including Grad-CAM, guided Grad-CAM, deconvolution visualization, and guided backpropagation, and interviewed mechanical workers on Amazon, demonstrating the superior performance of guided Grad-CAM, given its resemblance to human perception. However, this type of interpretability may be fragile. For example, we cropped or covered (with white polygons) the region with high activation in the original images. Nevertheless, the model still correctly recognizes most of the specimens, although with a low probability. By contrast, the activation patterns of Grad-CAM and guided Grad-CAM change and become difficult for humans to interpret. Regions with low activation in normal images were activated for accurate identification, indicating that class discrimination is not unique or immutable. In another study, neurons in CNNs have been confused by similar structures. In ImageNet, dog fur and wooden spoons, which have similar texture and color, have activated the same neurons (Olah et al. 2017). Similarly, the red stitches in baseballs have been confused with the white teeth and pink inner mouth of sharks (Carter et al. 2019). We found similar situations in images from the collected FID, as discussed earlier.

The 50 clades of fossils or fragments were successfully clustered (Fig. 9) using t-SNE. A total of 2000 test images exhibited 0.88 accuracy. Among them, the cluster of fossils with more affinity (or morphological) relationships was closer. Several superclusters were obtained, such as plants (classes 3, 22, and 35), vertebrates (classes 2, 4, 16, 26, 24, 30, 36, 39, 46, and 49), arthropods (classes 11, 23, and 27), fishes (classes 0, 12, 31, and 34), microfossils (classes 13, 19, 32, 37, and 43), teeth and bone fragments (classes 8, 25, 38, and 40), bivalves and brachiopods (classes 6 and 9), and the Asterozoa of starfish and ophiuroids (classes 29 and 44). We found that even for samples that seem to be clustered incorrectly, the predictions are not completely wrong. This condition may be critical, because CNNs have been used to quantitatively visualize the morphological characteristics of fossils or organisms (Esteva et al. 2017; Cuthill et al. 2019; Esgario et al. 2020; Liu and Song 2020). This case is not valid for a few specimens, such as the *Nautilus* specimen (class of 28) shown in Figure 9. The conical or cylindrical shell is similar to belemnite, but it is recognized correctly. This result suggests that flattening 3D feature maps into a 1D vector leads to the loss of spatial information. Moreover, the fully connected layer at the end of the DCNN architecture is important for the final classification. We also applied UMAP (McInnes et al. 2018) for dimensionality reduction and obtained similar results (Supplementary Fig. S3).

Considerable progress has been made in feature visualization over the past few years. The visual explanation provides a new perspective for understanding the workings of CNNs. However, the corresponding methods provide limited neuron interactions in CNNs (Olah et al. 2017), especially regarding quantitative visualization or morphological measurements.

Automatic Identification in Taxonomy.— Recent studies on deep learning for genus- or

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

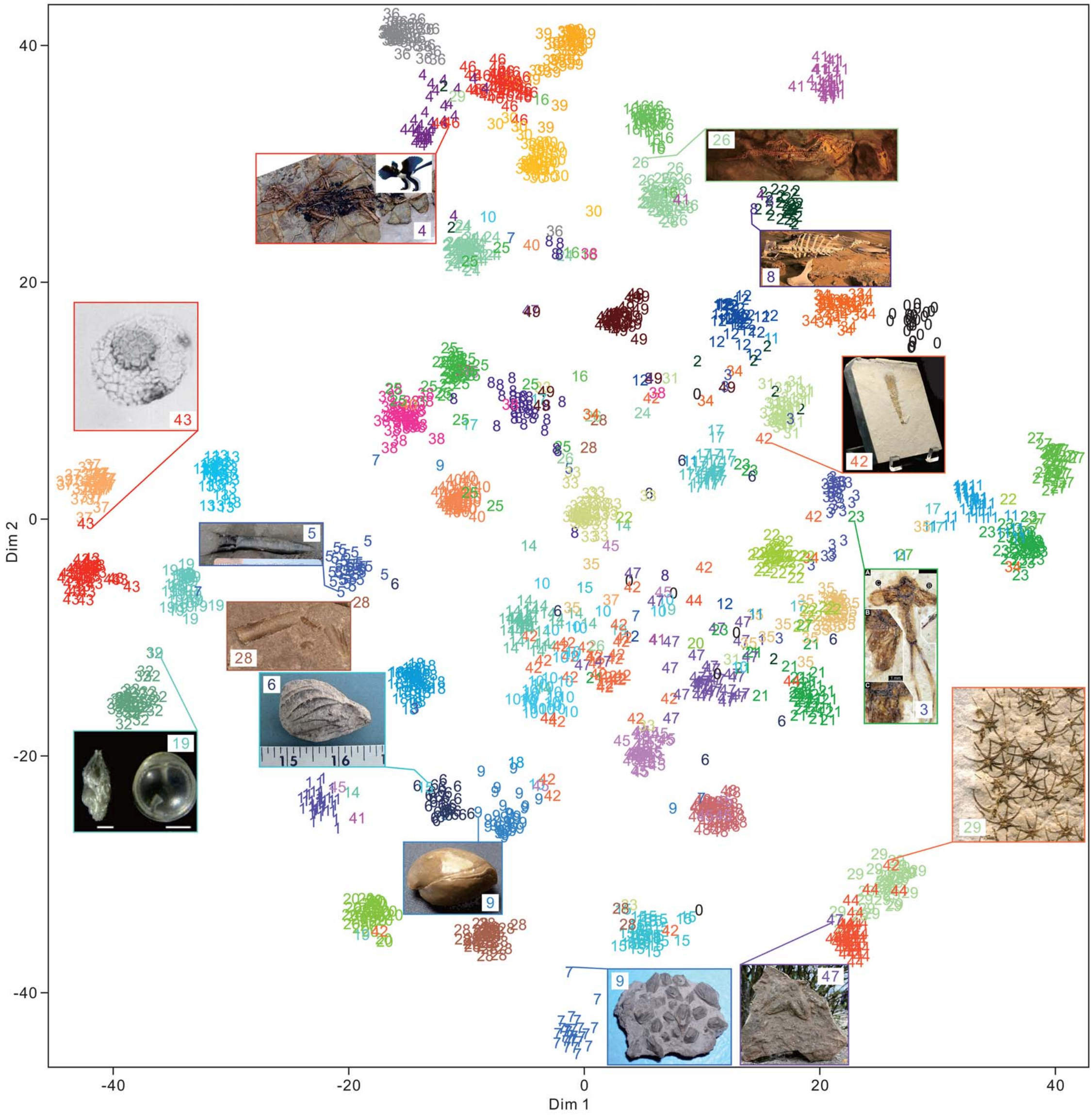

FIGURE 9. Feature visualization of the feature maps extracted from the final global average pooling layer of the Inception-ResNet-v2 architecture with 2000 random images (each class contains 40 images with 0.88 accuracy) in the test set using t-distributed stochastic neighbor embedding (t-SNE). The class order in alphabetical order is shown in Table 1. Some of the samples of clustering into other groups are shown in the rectangle with input images and their predicted labels. Specimens are not to scale. The image URLs are provided in Supplementary Table S1.

species-level identification have mainly focused on modern organisms and a few microfossils, giving generally scarce data. Hence, existing studies have covered dozens of common species in a particular geological period (Keçeli et al. 2017; Pires de Lima et al. 2020). We expanded this type of research and identified 50 fossil (fragment) clades, rather than focusing on the identification of several species, and the identification performance seems comparable to that of human experts. We believe that under data scarcity, automatic taxonomic identification can be gradually enhanced from high-level identification toward genus-/species-level identification of specific groups. For example, if the FID is further labeled at

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

the order, family, or genus level, then supervised taxonomic identification can be more detailed. However, image-based identification of fossil specimens has inherent limitations for some fossil groups. In fact, systematic taxonomists identify fossils not only by visual characteristics but also by considering additional structures, wall structures, and shell compositions of foraminifera, sutures in cephalopods, internal structures of brachiopods, and vein structures of leaves. Color images generally fail to reflect these features, making it difficult for DCNNs to detect them too. In our FID, images of sponges, bryozoans, trace fossils, and other clades showing various morphologies or with fewer training samples available led to inferior performance, indicating the difficulty of learning characteristics for these fossils compared with other clades. Pires de Lima et al. (2020) also demonstrated that machine classifiers consistently misidentify some of the fusulinid specimens, because the wall structure is neglected. Piazza et al. (2021) demonstrated a solution that used scanning electron microscopy images paired with a morphological matrix to recognize marine coralline algae. The matrix is connected with the flattened final feature maps and uses a fully connected layer for classification (Fig. 3). Thus, the chosen matrix is human intervention, which can add biotic or even abiotic information, but it loses the convenience of fully automatic implementation. In addition, to achieve species-level identification for certain fossil groups, other aspects should be considered, such as damaged specimens (Bourel et al. 2020) and the directions of thin-section specimens (Pires de Lima et al. 2020). Furthermore, secondary revision of the data collected from the literature is essential for the accurate supervision of deep learning (Hsiang et al. 2019). The taxonomic positions of some taxa may be modified in subsequent studies, it being necessary to coordinate classification criteria among scholars and research communities (Al-Sabouni et al. 2018; Fenton et al. 2018) for consistency. Even internal variations of taxa in different regions and periods should be considered, and DCNNs can also be used to verify these problems (Pires de Lima et al. 2020), given their highly accurate, reproducible, and unbiased classification (Renaudie et al. 2018; Hsiang et al. 2019; Marchant et al. 2020).

Deep learning may bring systematic paleontology to a new stage, and morphology-based manual taxonomic identification, including identification of microfossils and some invertebrate fossils, may soon be replaced by deep learning (Valan et al. 2019). Experiments on invertebrate specimens demonstrate that the performance of deep learning is comparable to that of taxonomists (Hsiang et al. 2019; Mitra et al. 2019). With the continuous digitization of geological data, more fossil clades will be included and performance will improve. Although DCNNs cannot identify new species (supervised learning can only identify existing fossils, whereas unsupervised learning may detect anomalies or new species), they can accurately solve routine and labor-intensive tasks without the prior knowledge that taxonomists can only acquire after several years of training. Machine learning classifiers can help experts devote their time to the most challenging and ambiguous identification cases (Romero et al. 2020). Recently, a single model was developed to identify thousands of common living plant species (Joly et al. 2016), which indicates that deep learning also has great potential in automatic taxonomy identifications. For our DCNN model to be publicly available, we deployed it on a web server, an uncommon practice in paleontology to date. Users can visit and use it at www.ai-fossil.com. The application of deep learning in paleontology benefits not only the academic community but also paleontological fieldwork, education, and museum collection management, spreading knowledge to the public. Furthermore, these applications can also provide more data for deep learning, and lead to more robust and accurate models.

Solutions to Data Scarcity.—To our knowledge, this is the first time that web crawlers have been used or reported as being used in paleontology to collect image data for applications to automatic fossil identification based on DCNNs. This data-collection approach provides a new opportunity for disciplines in which massive training sets are difficult to obtain for developing deep learning. The

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

hardware bottleneck of deep learning has mostly been overcome due to the dramatic increase in computing power of the graphics processor units, tensor processing units, and other artificial intelligence chips. However, the lack of large datasets poses a major obstacle for the application of deep learning in fields such as paleontology. We shared the FID, which can be reused in the future to train more powerful models and provide training data for genus/species identifications. Meanwhile, the trained model can be used for the rough identification of newly collected raw data, considering that data clearing/labeling is usually time-consuming.

With the increasing digitization and sharing of huge quantities of samples that have been housed in universities, research institutes, and museums and collected over the last three centuries, deep learning seems to be a promising research direction, and related efforts are underway, such as the Endless Forams (Hsiang et al. 2019) and GB3D Fossil Types Online Database (http://www.3d-fossils.ac.uk). Furthermore, the information age allows the use of diverse data sources through approaches such as citizen science data (Catlin-Groves 2012), which have been widely applied in biology through developments such as iNaturalist (Nugent 2018), e-Bird (Sullivan et al. 2009), eButterfly (Prudic et al. 2017), and Zooniverse (Simpson et al. 2014). In addition, public data from the Internet could be considered. Alternatively, various algorithms reduce the need for massive sets of labeled data by adopting approaches such as unsupervised learning (Caron et al. 2018), semi-supervised learning (Kipf and Welling 2017), and exploring deep learning on small datasets (Liu and Deng 2015), likely accelerating the application of deep learning in paleontology.

Conclusions

In this study, we used web crawlers to collect the FID from the Internet to alleviate the data deficiency for fossil clade identification. The FID contains 415,339 images belonging to 50 fossil clades that can be used to train and evaluate three DCNN architectures. The Inception-ResNet-v2 architecture achieved 0.90 top-1 accuracy and 0.97 top-3 accuracy in the test set. We also demonstrated that transfer learning is not only applicable to small datasets but is also very powerful and efficient when applied to large data volumes (~10^6 samples). We conducted visual explanation methods to reveal discriminative features and regions for taxonomic identification using deep learning. The results revealed similarities between taxonomists and algorithms in learning and performing fossil image identification. Data scarcity has become a major obstacle to the application of deep learning in paleontology. With the digitization of dark data (i.e., unstructured data) and the collection of data from multiple sources, image-based systematic taxonomy may soon be replaced by deep learning solutions. Furthermore, we contributed a website for the entire community to access the models. We deployed the DCNN on a server for end-to-end fossil clade identification at www.ai-fossil.com.

Acknowledgments

We thank X. Dai for identifying cephalopods and for preparing the FID. We thank GitHub users for sharing the web crawler packages that we used to prepare the FID, including google-images-download (https://github.com/hardikvasa/google-images-download) from H. Vasa et al. and Bing image downloader (https://github.com/ostrolucky/Bulk-Bing-Image-downloader) from G. Ostrolucký et al. We thank all the web pages and sites that contribute to the FID. We declare that all the collected images are used for academic purposes only. We thank two anonymous reviewers for their insightful comments. This study is supported by the National Natural Science Foundation of China (41821001) and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB26000000). This is the Center for Computational & Modeling Geosciences Publication Number 5.

Data Availability Statement

Data available from the Dryad and Zenodo digital repositories: https://doi.org/10.5061/dryad.51c59zwb0, https://doi.org/10.5281/zenodo.6333970.

Literature Cited

Abadi, M., P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, and M. Isard. 2016. Tensorflow: a system for large-scale machine learning. Pp. 265–283 *in* 12th USENIX Symposium on Operating Systems Design and Implementation. USENIX Association, Savannah, Ga.

Al-Barazanchi, H. A., A. Verma, and S. Wang. 2015. Performance evaluation of hybrid CNN for SIPPER plankton image calssification. Pp. 551–556 *in* Proceedings of 2015 Third International Conference on Image Information Processing (ICIIP). IEEE, Waknaghat, India.

Al-Barazanchi, H., A. Verma, and S. X. Wang. 2018. Intelligent plankton image classification with deep learning. International Journal of Computational Vision and Robotics 8:561–571.

Al-Sabouni, N., I. S. Fenton, R. J. Telford, and M. Kucera. 2018. Reproducibility of species recognition in modern planktonic foraminifera and its implications for analyses of community structure. Journal of Micropalaeontology 37:519–534.

Babenko, A., A. Slesarev, A. Chigorin, and V. Lempitsky. 2014. Neural codes for image retrieval. Pp. 584–599 *in* European Conference on Computer Vision. Springer, Zurich.

Beaufort, L., and D. Dollfus. 2004. Automatic recognition of coccoliths by dynamical neural networks. Marine Micropaleontology 51:57–73.

Botev, Z. I., D. P. Kroese, R. Y. Rubinstein, and P. L'Ecuyer. 2013. The cross-entropy method for optimization. Pp. 35–59 *in* C. R. Rao and V. Govindaraju, eds. Handbook of statistics. Elsevier, Amsterdam.

Bourel, B., R. Marchant, T. de Garidel-Thoron, M. Tetard, D. Barboni, Y. Gally, and L. Beaufort. 2020. Automated recognition by multiple convolutional neural networks of modern, fossil, intact and damaged pollen grains. Computers and Geosciences 140:104498.

Brodzicki, A., M. Piekarski, D. Kucharski, J. Jaworek-Korjakowska, and M. Gorgon. 2020. Transfer learning methods as a new approach in computer vision tasks with small datasets. Foundations of Computing and Decision Sciences 45:179–193.

Bueno, G., O. Deniz, A. Pedraza, J. Ruiz-Santaquiteria, J. Salido, G. Cristóbal, M. Borrego-Ramos, and S. Blanco. 2017. Automated diatom classification (part A): handcrafted feature approaches. Applied Sciences 7:753.

Byeon, W., M. Domínguez-Rodrigo, G. Arampatzis, E. Baquedano, J. Yravedra, M. A. Maté-González, and P. Koumoutsakos. 2019. Automated identification and deep classification of cut marks on bones and its paleoanthropological implications. Journal of Computational Science 32:36–43.

Caron, M., P. Bojanowski, A. Joulin, and M. Douze. 2018. Deep clustering for unsupervised learning of visual features. Pp. 132–149 *in* Proceedings of the European Conference on Computer Vision (ECCV). Springer, Munich.

Carter, S., Z. Armstrong, L. Schubert, I. Johnson, and C. Olah. 2019. Activation atlas. Distill 4:e15.

Carvalho, L., G. Fauth, S. B. Fauth, G. Krahl, A. Moreira, C. Fernandes, and A. Von Wangenheim. 2020. Automated microfossil identification and segmentation using a deep learning approach. Marine Mcropaleontology 158:101890.

Catlin-Groves, C. L. 2012. The citizen science landscape: from volunteers to citizen sensors and beyond. International Journal of Zoology 2012:349630.

Chollet, F. 2015. Keras. https://github.com/fchollet/keras, accessed 13 August 2021.

Cuthill, J. F. H., N. Guttenberg, S. Ledger, R. Crowther, and B. Huertas. 2019. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Science Advances 5: eaaw4967.

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. 2009. Imagenet: a large-scale hierarchical image database. Pp. 248–255 *in* 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Miami.

Domínguez-Rodrigo, M., and E. Baquedano. 2018. Distinguishing butchery cut marks from crocodile bite marks through machine learning methods. Scientific Reports 8:5786.

Esgario, J. G., R. A. Krohling, and J. A. Ventura. 2020. Deep learning for classification and severity estimation of coffee leaf biotic stress. Computers and Electronics in Agriculture 169:105162.

Esteva, A., B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118.

Fawcett, T. 2006. An introduction to ROC analysis. Pattern Recognition Letters 27:861–874.

Fenton, I. S., U. Baranowski, F. Boscolo-Galazzo, H. Cheales, L. Fox, D. J. King, C. Larkin, M. Latas, D. Liebrand, and C. G. Miller. 2018. Factors affecting consistency and accuracy in identifying modern macroperforate planktonic foraminifera. Journal of Micropalaeontology 37:431–443.

Fukui, H., T. Hirakawa, T. Yamashita, and H. Fujiyoshi. 2019. Attention branch network: learning of attention mechanism for visual explanation. Pp. 10705–10714 in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, Long Beach, Calif.

Fukushima, K. 1980. Neocognitron: a self-organizing neural network model for a mechanism of visual pattern recognition. Biological Cybernetics 36:193–202.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21:1263–1284.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Identity mappings in deep residual networks. Pp. 630–645 *in* European Conference on Computer Vision. Springer, Amsterdam.

Helfenstein, A., and P. Tammela. 2017. Analyzing user-generated online content for drug discovery: development and use of MedCrawler. Bioinformatics 33:1205–1209.

Hinton, G. E., and R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. Science 313:504–507.

Hinz, T., N. Navarro-Guerrero, S. Magg, and S. Wermter. 2018. Speeding up the hyperparameter optimization of deep convolutional neural networks. International Journal of Computational Intelligence and Applications 17:1850008.

Hou, Y., X. Cui, M. Canul-Ku, S. Jin, R. Hasimoto-Beltran, Q. Guo, and M. Zhu. 2020. ADMorph: a 3D digital microfossil morphology dataset for deep learning. IEEE Access 8:148744–148756.

Hsiang, A. Y., A. Brombacher, M. C. Rillo, M. J. Mleneck-Vautravers, S. Conn, S. Lordsmith, A. Jentzen, M. J. Henehan, B. Metcalfe, and I. S. Fenton. 2019. Endless forams: >34,000 modern planktonic foraminiferal images for taxonomic training and automated species recognition using convolutional neural networks. Paleoceanography and Paleoclimatology 34:1157–1177.

Hubel, D. H., and T. N. Wiesel. 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. Journal of Pysiology 160:106–154.

Joly, A., P. Bonnet, H. Goëau, J. Barbe, S. Selmi, J. Champ, S. Dufour-Kowalski, A. Affouard, J. Carré, and J.-F. Molino. 2016. A look inside the Pl@ntNet experience. Multimedia Systems 22:751–766.

Kausar, M. A., V. Dhaka, and S. K. Singh. 2013. Web crawler: a review. International Journal of Computer Applications 63:31–36.

Kaya, A., A. S. Keceli, C. Catal, H. Y. Yalic, H. Temucin, and B. Tekinerdogan. 2019. Analysis of transfer learning for deep neural network based plant classification models. Computers and Electronics in Agriculture 158:20–29.

Keçeli, A. S., A. Kaya, and S. U. Keçeli. 2017. Classification of radiolarian images with hand-crafted and deep features. Computers and Geosciences 109:67–74.

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

Keçeli, A. S., S. U. Keçeli, and A. Kaya. 2018. Classification of radiolarian fossil images with deep learning methods. Pp. 1–4 *in* 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, Izmir, Turkey.

Kipf, T. N., and M. Welling. 2017. Semi-supervised classification with graph convolutional networks. Pp. 1–14 *in* Fifth International Conference on Learning Representations. IEEE, Toulon, France.

Kloster, M., D. Langenkämper, M. Zurowietz, B. Beszteri, and T. W. Nattkemper. 2020. Deep learning-based diatom taxonomy on virtual slides. Scientific Reports 10:14416.

Koeshidayatullah, A., M. Morsilli, D. J. Lehrmann, K. Al-Ramadan, and J. L. Payne. 2020. Fully automated carbonate petrography using deep convolutional neural networks. Marine and Petroleum Geology 122:104687.

Kong, S., S. Punyasena, and C. Fowlkes. 2016. Spatially aware dictionary learning and coding for fossil pollen identification. Pp. 1305–1314 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, Las Vegas, Nev.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. Pp. 1097–1105 *in* Neural Information Processing Systems Conference and Workshop. Curran Associates, Lake Tahoe, Calif.

Lambert, D., and R. Green. 2020. Automatic identification of diatom morphology using deep learning. Pp. 1–7 *in* 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ). IEEE, Wellington, New Zealand.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2278–2324.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature 521:436–444.

Li, H., Z. Lin, X. Shen, J. Brandt, and G. Hua. 2015. A convolutional neural network cascade for face detection. Pp. 5325–5334 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Boston.

Li, X., and Z. Cui. 2016. Deep residual networks for plankton classification. Pp. 1–4 *in* OCEANS 2016 MTS/IEEE Monterey. IEEE, Monterey, Calif.

Liu, B., Y. Zhang, D. He, and Y. Li. 2018a. Identification of apple leaf diseases based on deep convolutional neural networks. Symmetry 10:11.

Liu, C., B. Zoph, M. Neumann, J. Shlens, W. Hua, L.-J. Li, L. Fei-Fei, A. Yuille, J. Huang, and K. Murphy. 2018b. Progressive neural architecture search. Pp. 19–34 *in* European Conference on Computer Vision. Springer, Munich.

Liu, S., and W. Deng. 2015. Very deep convolutional neural network based image classification using small training sample size. Pp. 730–734 *in* 2015 Third IAPR Asian Conference on Pattern Recognition (ACPR). IEEE, Kuala Lumpur, Malaysia.

Liu, X., and H. Song. 2020. Automatic identification of fossils and abiotic grains during carbonate microfacies analysis using deep convolutional neural networks. Sedimentary Geology 410:105790.

Long, J., E. Shelhamer, and T. Darrell. 2015. Fully convolutional networks for semantic segmentation. Pp. 3431–3440 *in* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Boston.

Lopez-Aparicio, S., H. Grythe, M. Vogt, M. Pierce, and I. Vallejo. 2018. Webcrawling and machine learning as a new approach for the spatial distribution of atmospheric emissions. PLoS ONE 13: e0200650.

Maaten, L. v. d., and G. Hinton. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9:2579–2605.

MacLeod, N. 2007. Automated taxon identification in systematics: theory, approaches and applications. CRC Press, Boca Raton, FL.

MacLeod, N., and L. Kolska Horwitz. 2020. Machine-learning strategies for testing patterns of morphological variation in small samples: sexual dimorphism in gray wolf (*Canis lupus*) crania. BMC Biology 18:113.

MacLeod, N., M. Benfield, and P. Culverhouse. 2010. Time to automate identification. Nature 467:154–155.

Marchant, R., M. Tetard, A. Pratiwi, M. Adebayo, and T. de Garidel-Thoron. 2020. Automated analysis of foraminifera fossil records by image classification using a convolutional neural network. Journal of Micropalaeontology 39:183–202.

Marcos, J. V., R. Nava, G. Cristóbal, R. Redondo, B. Escalante-Ramírez, G. Bueno, Ó. Déniz, A. González-Porto, C. Pardo, and F. Chung. 2015. Automated pollen identification using microscopic imaging and texture analysis. Micron 68:36–46.

Martineau, M., D. Conte, R. Raveaux, I. Arnault, D. Munier, and G. Venturini. 2017. A survey on image-based insect classification. Pattern Recognition 65:273–284.

McInnes, L., J. Healy, and J. Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML].

Mitra, R., T. Marchitto, Q. Ge, B. Zhong, B. Kanakiya, M. Cook, J. Fehrenbacher, J. Ortiz, A. Tripati, and E. Lobaton. 2019. Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. Marine Micropaleontology 147:16–24.

Ngugi, L. C., M. Abelwahab, and M. Abo-Zahhad. 2021. Recent advances in image processing techniques for automated leaf pest and disease recognition—a review. Information Processing in Agriculture 8:27–51.

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences USA 115:E5716–E5725.

Nugent, J. 2018. iNaturalist. Science Scope 41:12–13.

Olah, C., A. Mordvintsev, and L. Schubert. 2017. Feature visualization. Distill 2:e7.

Olah, C., A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev. 2018. The building blocks of interpretability. Distill 3:e10.

Pankhurst, R. J. 1974. Automated identification in systematics. Taxon 23:45–51.

Pedraza, A., G. Bueno, O. Deniz, G. Cristóbal, S. Blanco, and M. Borrego-Ramos. 2017. Automated diatom classification (part B): a deep learning approach. Applied Sciences 7:460.

Piazza, G., C. Valsecchi, and G. Sottocornola. 2021. Deep learning applied to SEM images for supporting marine coralline algae classification. Diversity 13:640.

Pires de Lima, R., K. F. Welch, J. E. Barrick, K. J. Marfurt, R. Burkhalter, M. Cassel, and G. S. Soreghan. 2020. Convolutional neural networks as an aid to biostratigraphy and micropaleontology: a test on Late Paleozoic microfossils. Palaios 35:391–402.

Prudic, K. L., K. P. McFarland, J. C. Oliver, R. A. Hutchinson, E. C. Long, J. T. Kerr, and M. Larrivée. 2017. eButterfly: leveraging massive online citizen science for butterfly conservation. Insects 8:53.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi. 2016. You only look once: Unified, real-time object detection. Pp. 779–788 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE, Las Vegas, NV.

Renaudie, J., R. Gray, and D. B. Lazarus. 2018. Accuracy of a neural net classification of closely-related species of microfossils from a sparse dataset of unedited images. PeerJ 6:e27328v1.

Rodner, E., M. Simon, G. Brehm, S. Pietsch, J. W. Wägele, and J. Denzler. 2015. Fine-grained recognition datasets for biodiversity analysis. Pp. 1–4 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Boston.

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press

Romero, I. C., S. Kong, C. C. Fowlkes, C. Jaramillo, M. A. Urban, F. Oboh-Ikuenobe, C. D'Apolito, and S. W. Punyasena. 2020. Improving the taxonomy of fossil pollen using convolutional neural networks and superresolution microscopy. Proceedings of the National Academy of Sciences USA 117:28496–28505.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. 1986. Learning representations by back-propagating errors. Nature 323:533–536.

Selvaraju, R. R., A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra. 2016. Grad-CAM: why did you say that? arXiv:1611.07450 [stat.ML].

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. Pp. 618–626 *in* IEEE International Conference on Computer Vision. IEEE, Venice, Italy.

Sevillano, V., and J. Aznarte. 2018. Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. PLoS ONE 13:e0201807.

Shorten, C., and T. M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. Journal of Big Data 6:1–48.

Simonyan, K., and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV].

Simonyan, K., A. Vedaldi, and A. Zisserman. 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:1312.4400 [cs.CV].

Simpson, R., K. R. Page, and D. De Roure. 2014. Zooniverse: observing the world's largest citizen science platform. Pp. 1049–1054 *in* Proceedings of the 23rd International Conference on World Wide Web. Association for Computing Machinery, New York.

Sokolova, M., and G. Lapalme. 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45:427–437.

Sullivan, B. L., C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling. 2009. eBird: a citizen-based bird observation network in the biological sciences. Biological Conservation 142:2282–2292.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2015. Going deeper with convolutions. Pp. 1–9 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Boston.

Szegedy, C., S. Ioffe, V. Vanhoucke, and A. Alemi. 2017. Inception-v4, Inception-ResNet and the impact of residual connections on learning. Pp. 4278–4284 *in* 31st AAAI Conference on Artificial Intelligence. AAAI, San Francisco.

Tabak, M. A., M. S. Norouzzadeh, D. W. Wolfson, S. J. Sweeney, K. C. VerCauteren, N. P. Snow, J. M. Halseth, P. A. Di Salvo, J. S. Lewis, and M. D. White. 2019. Machine learning to classify animal species in camera trap images: applications in ecology. Methods in Ecology and Evolution 10:585–590.

Tan, C., F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. 2018. A survey on deep transfer learning. Pp. 270–279 *in* International Conference on Artificial Neural Networks. Springer, Rhodes, Greece.

Tetard, M., R. Marchant, G. Cortese, Y. Gally, T. de Garidel-Thoron, and L. Beaufort. 2020. A new automated radiolarian image acquisition, stacking, processing, segmentation and identification workflow. Climate of the Past 16:2415–2429.

Too, E. C., L. Yujian, S. Njuki, and L. Yingchun. 2019. A comparative study of fine-tuning deep learning models for plant disease identification. Computers and Electronics in Agriculture 161:272–279.

Urbankova, P., V. Scharfen, and J. Kulichová. 2016. Molecular and automated identification of the diatom genus *Frustulia* in northern Europe. Diatom Research 31:217–229.

Valan, M., K. Makonyi, A. Maki, D. Vondráček, and F. Ronquist. 2019. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. Systematic Bology 68:876–895.

Villa, A. G., A. Salazar, and F. Vargas. 2017. Towards automatic wild animal monitoring: identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological Informatics 41:24–32.

Wang, C., R. M. Hazen, Q. Cheng, M. H. Stephenson, C. Zhou, P. Fox, S.-z. Shen, R. Oberhänsli, Z. Hou, X. Ma, Z. Feng, J. Fan, C. Ma, X. Hu, B. Luo, J. Wang, and C. M. Schiffries. 2021. The Deep-Time Digital Earth program: data-driven discovery in geosciences. National Science Review 7:nwab027.

Wang, J., and L. Perez. 2017. The effectiveness of data augmentation in image classification using deep learning. arXiv:1712.04621 [cs.CV].

Wang, Y.-X., D. Ramanan, and M. Hebert. 2017. Growing a brain: fine-tuning by increasing model capacity. Pp. 2471–2480 *in* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, Honolulu.

Willi, M., R. T. Pitman, A. W. Cardoso, C. Locke, A. Swanson, A. Boyer, M. Veldthuis, and L. Fortson. 2019. Identifying animal species in camera trap images using deep learning and citizen science. Methods in Ecology and Evolution 10:80–91.

Xiao, T., T. Xia, Y. Yang, C. Huang, and X. Wang. 2015. Learning from massive noisy labeled data for image classification. Pp. 2691–2699 *in* Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. IEEE, Boston.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. How transferable are features in deep neural networks? Pp. 3320–3328 *in* Proceedings of the 27th International Conference on Neural Information Processing Systems. MIT, Cambridge, Mass.

Zeiler, M. D., and R. Fergus. 2014. Visualizing and Understanding Convolutional Networks. Pp. 818–833 *in* European Conference on Computer Vision. Springer, Zurich.

Zeiler, M. D., G. W. Taylor, and R. Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. Pp. 2018–2025 *in* 2011 International Conference on Computer Vision. IEEE, Barcelona.

Zhong, B., Q. Ge, B. Kanakiya, R. M. T. Marchitto, and E. Lobaton. 2017. A comparative study of image classification algorithms for Foraminifera identification. Pp. 1–8 *in* 2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, Honolulu.

https://doi.org/10.1017/pab.2022.14 Published online by Cambridge University Press