Egypt Monuments Dataset version 1: A Scalable Benchmark for Image Classification and Monument Recognition

Mennat Allah Hassan[1,2,*], Alaa Hamdy[1], Mona Nasr[2]
Faculty of Computer Science, Misr International University, Cairo, Egypt[1]
Faculty of Computers and Artificial Intelligence, Helwan University, Cairo, Egypt[2]

***Abstract*—The success of machine learning (ML) as well as deep learning (DL) depends largely on data availability and quality. The system's performance is frequently more affected by the amount and quality of its training data than by its architecture and training specifics. Consequently, demand exists for challenging datasets that both precisely measure performance and present unique challenges with real-world applications. The Egypt Monuments Dataset v1 (EGYPT-v1) is introduced as a new scalable benchmark for fine-image classification (IC) and object recognition (OR) in the domain of ancient Egyptian monuments. EGYPT-v1 dataset is by far the world's first large specified such dataset to date, with over seven thousand images and 40 distinct instance labels. The dataset composes different categories of monuments such as pyramids, temples, mummies, statues, head statues, bust statues, heritage sites, palaces and shrines. Several advanced deep network architectures were tested to appraise the classification difficulty in the EGYPT-v1 dataset, namely ResNet50, Inception V3, and LeNet5 models. The models achieved accuracy rates as follows: 99.13%, 90.90%, and 92.64%, respectively. The dataset was predominantly created by manually collecting images from the popular global online video-sharing and social media platform, Youtube, as well as WATCHiT, Egypt's top streaming entertainment service. Additionally, Wikimedia Commons, the largest crowdsourced media repository in the world, was used as a secondary source of images. The images that comprise the dataset can be accessed on the GitHub repository https://github.com/mennatallahhassan/egypt-monuments-dataset.**

***Keywords*—Deep learning; landmark datasets; landmark recognition; monument datasets; monument recognition**

I. Introduction

Within the realm of computer vision, IC and OR are key research topics that have been extensively investigated for years. The objective of IC is to [1], [2], [3], [4]. The objective of OR [5], [6] is the computer vision task of recognising a particular instance of an object, as opposed to its category. For example, it is interested in instance-level labels such as "Karnak Temple in Luxor" or "Khufu Pyramid in Giza" rather than simply "Karnak" or "Khufu" when labeling images.

When ML and DL techniques for image classification and instance-level recognition (ILR) tasks have progressed, methods have improved in their robustness and scalability, and they have started solving standard datasets.

Furthermore, despite the fact that increasingly largescale classification datasets, for instance, CIFAR-10 [7], ImageNet [8], in addition to OpenImages [9], have become standard benchmarks, there still needs to be monument datasets for fine-grained instance recognition and classification. A monument refers to a structure that has been erected and can take many forms, including busts, crosses, statues, fountains, mausoleums, obelisks, pyramids, reliquaries, sarcophagi, steles, graves, or triumphal arches. In addition, smaller-scale forms such as medals and commemorative plaques can also be considered monuments [10]. Generally, the world is full of monuments; remarkably, Egypt contains a third of the world's monuments; parts of these monuments are displayed in the most famous museums all over the world [11]. This paper presents the Egypt Monuments Dataset v1 (EGYPT-v1), a novel scalable dataset for IC and ILR. More than seven thousand images of forty-one different monuments and historical landmarks are available in EGYPT-v1, as seen in Fig. 1. Fig. 2 illustrates its geographic distribution across Egypt. The instance recognition task uses a training dataset of 5,833 labelled images and 1,945 labelled images as a test set, that includes ground truth information regarding the instance classification (IC) and ILR tasks. Although our primary objective for Egypt Monuments Dataset v1 is to recognize historical landmarks and monuments, the solutions developed to overcome the obstacles, can be easily adapted to address other instance-level recognition challenges, including artwork recognition.

The primary objective of Egypt Monuments Dataset v1 is to replicate real-world circumstances and, consequently, introduces several complex hindrances. There are thousands of images representing tens of classes in EGYPT-v1. The degree of intra-class variation is significantly elevated, with images of a single class exhibiting both indoor and outdoor views and images that possess a tangential association with a particular class, like museum paintings. The Egypt Monuments Dataset v1 is intended to be used as a novel benchmark for IC and ILR.

The dataset, training instance labels, classification and recognition ground truth data, and metric computation code are accessible to the public.

In summary, this paper presents the Egypt Monuments Dataset v1, a novel and challenging benchmark for fine-image classification and object recognition in the domain of ancient Egyptian monuments. The RELATED WORK section compares the existing monument/landmark image classification and recognition datasets with our novel proposed dataset. The dataset consists of many images and distinct instance labels

Fig. 1. An overview of the EGYPT-v1 dataset composed of over 7k images for 41 classes.

TABLE I. THE EGYPT-V1 DATASET IS BEING HIGHLIGHTED AMONG EXISTING MONUMENT AND HERITAGE SITES DATASETS. THE EGYPT MONUMENTS DATASET V1 IS THE FIRST PUBLIC DATASET REGARDING THE AVAILABILITY OF EGYPTIAN ANCIENT MONUMENT IMAGES AND HERITAGE SITES

Dataset	Year	# Monuments/Landmarks	# Images	Annotation Collection	Dataset Scale
Oxford [1]	2007	11	5,063	Manual	City
Paris [12]	2008	11	6,392	Manual	City
Holidays [13]	2008	500	-	Manual	Worldwide
European Cities 50k [14]	2010	20	50k	Manual	Continent
Geotagged StreetView [15]	2010	-	17k	StreetView	City
Rome 16k [16]	2010	69	16k	GeoTag + SfM	City
San Francisco [17]	2011	-	1.7M	StreetView	City
Landmarks-PointCloud [18]	2012	1k	205k	Flickr label + SfM	Worldwide
Singapore Landmark-40 [19]	2012	40	13,538	Internet sources + Manual	City
24/7 Tokyo [20]	2015	125	1k	Smartphone + Manual	City
Paris500k [21]	2015	13k	501k	Manual	City
Landmark URLs [3]	2016	586	-	Text query + Feature matching	Worldwide
Google Landmarks [22]	2017	30k	1M	GPS + semi-automatic	Worldwide
Revisited Oxford [4]	2018	11	1M	Manual + semi-automatic	Worldwide
Revisited Paris [4]	2018	11	1M	Manual + semi-automatic	Worldwide
Qutub Complex Monuments' Images [23]	2018	5	1,286	Google Images	City
Indian heritage monuments [24]	2020	143	7,150	Web Scraping	Country
Google Landmarks Dataset v2 [25]	2019	200k	5M	Crowdsourced + semi-automatic	Worldwide
Our Egypt Monuments Dataset v1	2022	41	7,778	Manual + semi-automatic	Country

representing various categories of monuments. All details have been explained in the DATASET OVERVIEW section. The performance of several advanced deep network architectures on the EGYPT-v1 dataset has been evaluated. The accuracy rates have been illustrated in the EXPERIMENT section.

This paper also discusses the dataset's creation, distribution, and potential applications in instance-level recognition challenges, including recognition. Overall, the EGYPT-v1 dataset aims to replicate real-world circumstances and provide a valuable tool for researchers and practitioners in computer vision.

II. RELATED WORK

Image recognition's challenges extend from simple image classification (e.g., "human face" or "building") through fine-grained tasks that distinguish between models, and styles (such as "Head Sculpture" and "Ancient Temple") to recognition on an instance level (as "Portrait Head of Queen Tiye at the Neues Museum, Berlin, German" and "The Great Temple of Ramesses II, Aswan, southern Egypt"). Identifying ancient Egyptian monuments and historical landmarks is the primary focus of our novel dataset. Subsequently, datasets for image classification and recognition were scrutinized, with particular attention given to those most relevant to our research. Table I presents the existing datasets for monument/landmark image classification and recognition, along with our proposed novel dataset.

1) City-scale datasets: The datasets regarding Oxford [1], as well as Paris [12], consist of a huge number of landmark images found in both cities, collectively belonging to 11 categories. Additional datasets concentrate on photography from a specific city: Rome 16k [16]; Singapore Landmark-40 [19], including over 13,000 images from Singapore city. The dataset used in this study was created by collecting images from a range of different sources. Specifically, 40% of the images were obtained from Google Images, 40% were sourced from Flickr, and 5% came from Photobucket. The remaining 15% were acquired through manual means. This

portion comprised Geotagged Streetview Images [15], which consisted of approximately 17,000 photographs of Paris and San Francisco Landmarks [17], containing over 1.7 million images; Qutub Complex Monuments' Images containing 1,286 images for five famous monuments in Delhi, India; 24/7 Tokyo [20], including a thousand images under various lighting situations; and Paris500k [21], including 501,000 images.

2) Country-scale dataset: Indian heritage monuments dataset (IHMD) [24] contains 6,959 images of 413 classes. This dataset has been collected from image search engines using web scrappings such as Google Images, Bing Images, Wikimedia and Flickr.

3) Continent-scale datasets: The datasets that are more recent in origin contain images that have been sourced from a notably wider range of locales within the same continent than the older datasets. Within the European Cities (EC) 50k dataset, there are images of 20 landmarks that are distributed across 9 cities [14], including unannotated images from 5 additional cities that were used as distractors. Another version of this dataset has 1 million photos from 22 cities; however, all annotated photographs come from only one location [26].

4) Worldwide-scale datasets: The more expanded datasets have images of landmarks globally. The Landmarks-PointCloud dataset includes 205,000 images of 1,000 well-known landmarks [18]; The Landmark URLs dataset of approximately 192,000 images is classified into 586 landmarks. 168,882 images are utilized for fine-tuning in experiments, while the remaining 20,668 images are utilized to validate parameters [3]. The Revisited Oxford dataset, as well as Revisited Paris dataset are two other recent examples of global landmark datasets, each comprising eleven landmarks and roughly a million images [27]. The Google Landmarks Dataset, which is the dataset from which the data used in this study was drawn, originally consisted of 2.3 million photographs taken at 30,000 unique landmarks. However, this dataset is unstable due to copyright limitations. It declines over time as photographs are destroyed by the users who upload them [22]. To our best knowledge, no such large unique, collected country-scale datasets with ground truth Egyptian ancient monuments and landmarks visibility information are publicly available yet.

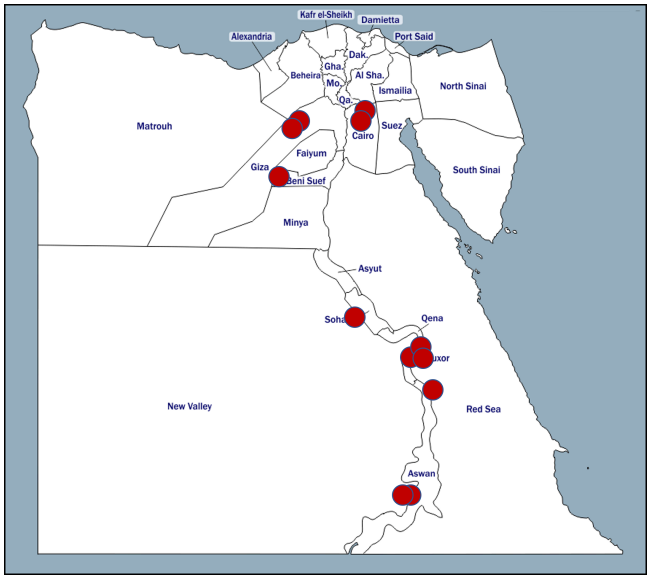


Fig. 2. Egypt map highlighting the distribution of ancient Egyptian monuments and heritage sites across Egypt regarding Egypt Monuments Dataset v1. (Starting from Cairo and reaching to Aswan).

III. Dataset Overview

A. Purposes

The purpose of the Egypt Monuments Dataset v1 is to simulate the following constraints of industrial monument/landmark recognition. It is *scalable* to encompass all ancient Egyptian monuments and landmarks worldwide, as they are not confined to Egypt alone. There are numerous discoveries and expeditions concerning ancient Egyptian monuments *Intra-class variability*. Images of monuments and historical landmarks are captured both indoors and outside, in a variety of lighting circumstances and from a variety of vantage points. In addition, there will be images that are connected to well-known ones. *Public availability.* The dataset aims to help the research community solve the scarcity of Egyptian ancient monument datasets that face researchers in this domain [28]. Our dataset was explicitly built to account for these difficulties.

B. Data Distribution

The Egypt Monuments Dataset v1 contains 41 diverse monuments and heritage sites from 6 out of the 28 governorates of Egypt; As indicated in Table II, this dataset of ancient Egyptian monuments and landmarks is truly unique and one-of-a-kind. By far, statutes are the most frequent type, followed by temples, then pyramids. Approximately 37% of the monuments with over 2,000 images are located in Luxor, while about 27% are in Cairo.

C. Dataset Construction

The process of gathering data and constructing the ground truth is described in this section.

1) Data sources: WatchiT, a leading Egyptian streaming entertainment platform, and Youtube, a global online video-sharing platform, are the primary sources for the Egypt Monuments Dataset v1. Then there is Wikimedia Commons, the most extensive online collection of user-submitted images, videos, and other media. Millions of photos of famous landmarks, taken by an active community of photographers and partner organizations like libraries, archives, and museums, are available on Wikimedia Commons under Creative Commons and Public Domain licenses. The goal of Wiki Loves Monuments, a yearly competition, is to add more high-quality landmark images to the site, while classifying them based on a detailed taxonomy of cultural heritage sites around the world. Images were also sourced from Google Images in addition to the aforementioned Wikimedia Commons.

2) Annotation: Notably, ground-truth annotation is notoriously difficult. Given that it is difficult to predefine what monuments or heritage sites are and that they are only sometimes clearly apparent, identifying monuments is challenging. Furthermore, for certain heritage sites, such as The Giza Pyramids and The Bent Pyramid, images can be captured from a considerable distance.

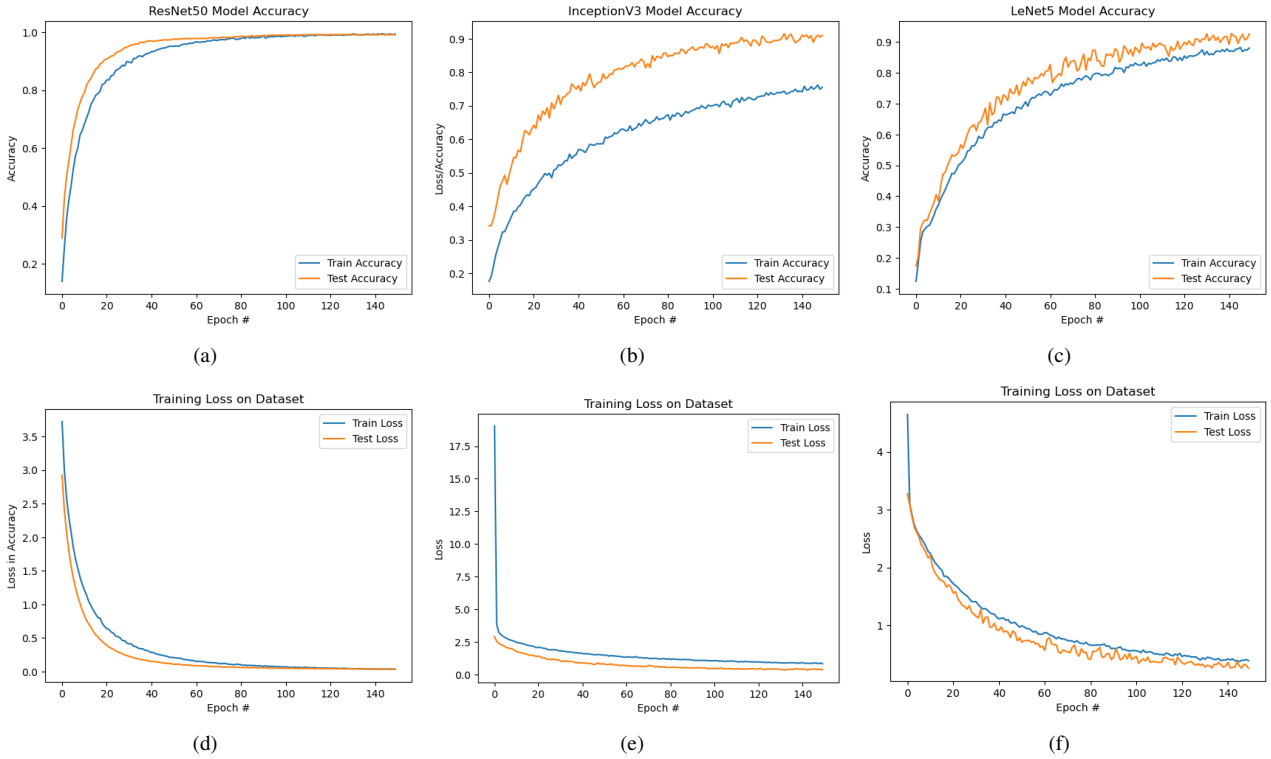


Fig. 3. (a) ResNet50 Model's Accuracy. (b) InceptionV3 Model's Accuracy. (c) LeNet5 Model's Accuracy. (d) Training Loss of ResNet50 Model. (e) Training Loss of InceptionV3 Model. (f) Training Loss of LeNet5 Model.

IV. Experiment

In this study, the utilization of the dataset is illustrated, and various baseline models that can be used as a reference for future research are introduced. Furthermore, an analysis of the outcomes obtained from the real-world challenge is provided. The findings discussed in this part are all relevant to the ground truth of version 1 of the dataset. Herein, the efficacy of recognizing the visual features of the EGYPT-v1 dataset using cutting-edge classification techniques is evaluated.

Experiments using various advanced deep network architectures have been conducted, such as ResNets [29], Inception V3 [30], and LeNet5 [31] models, to assess the level of classification difficulty in our EGYPT-v1 dataset. To train the models, data augmentation techniques have been employed.

Networks have been fine-tuned using pre-trained weights of ImageNet, optimized with Adam [32] and 1e-05 for the learning rate. Training and testing have been conducted with images sized at 224 × 224.

As shown in Table III, a performance comparison table was created to differentiate the three models. The models under examination are RESNET50, Inception V3, and LeNet 5, all of which are representative of deep learning techniques. The results of our analysis are presented in Table III in the document.

Upon close scrutiny of the graph presented in Fig. 3(a), it can be noted that the testing accuracy of ResNet50 surpasses the training accuracy after 90 epochs. In addition, a trend of improvement is observed in both the training and testing accuracy curves as the number of epochs progresses. The ResNet50 model attains 99.69% for training accuracy and 99.13% for testing accuracy at the completion of the training process. The dissimilarity between these two accuracy values is negligible, which indicates that the model is not partial to training images, and can perform with comparable efficiency in recognizing unobserved images. In contrast, as depicted in Fig. 3(d), the loss function performance is nearly indistinguishable for both curves. It is worth noting that saturation is observed for both curves at around epoch 110. The ability of the loss function to produce consistent results suggests that the model is not prone to overfitting concerns and can distinguish unknown data as efficiently as it classifies recognized data.

In Fig. 3(b), upon completion of training, InceptionV3's accuracy on the training data was 93.31%, while its accuracy on testing data was 90.90%. The substantial gap in accuracy scores indicates overfitting, a situation in which a model excels on training data while it shows inferior performance on new, unseen data. Overfitting can be addressed by reducing the model's complexity, increasing the training data volume, or implementing regularisation techniques. In Fig. 3(e), if the loss graph is examined closely, it can be seen that initially, the testing loss was significantly lower than the training loss. The training and testing loss demonstrated a decreasing trend as the epochs' number increased. Based on Fig. 3(f), it appears that the model has not fully converged and may benefit from additional training epochs. To confirm this speculation, an additional experiment was conducted, as shown in Fig. 4. The InceptionV3 model was trained on the same dataset with 600 epochs and achieved an accuracy of 96% with a loss curve

TABLE II. INSIGHTS OF THE EGYPT-V1 DATASET, CONTAINING 41 DIVERSE CLASSES DISTRIBUTED IN 6 GOVERNORATES OF EGYPT

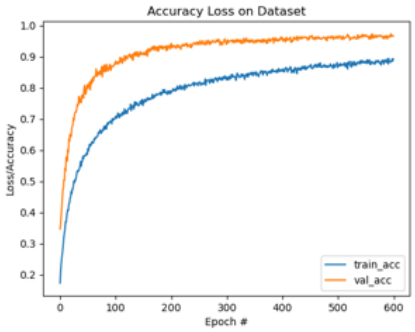
Monument/Heritage Site Name	Category	Located in	# Images
The Great Temple of Ramesses II	Temple	Aswan	1283
Hatshepsut Temple-Deir ElBahari Temple	Temple	Luxor	590
The Shunet El Zebib	Heritage Site	Sohag	558
Saqqara Pyramid-Pyramid of Djoser-Step Pyramid of Djoser	Pyramid	Giza	445
The Great Sphinx of Giza-Abou El Houl	Statue	Giza	358
Bent Pyramid of King Sneferu	Pyramid	Giza	305
The Small Temple of Abu Simbel-Temple of Nefertari-Temple of Hathor	Temple	Aswan	280
Menkaure Pyramid	Pyramid	Giza	267
Khafre Pyramid	Pyramid	Giza	247
Medinet Habu Temple-The Temple of Ramses III	Temple	Luxor	242
Meidum Pyramid of King Sneferu	Pyramid	Beni Suef	213
Head Statue of Akhenaten	Statue	Luxor	206
Karnak Temple	Temple	Luxor	205
Architect Senenmut with Princess Neferu-Ra	Statue	Cairo	193
Red Hatshepsut Shrines	Shrine	Luxor	166
Malkata Palace-Amenhotep III Palace	Palace	Luxor	152
Statue of King Zoser	Statue	Cairo	152
Mask of Tutankhamun	Mask	Cairo	147
King Amenhotemp Shrine	Shrine	Luxor	141
Abu Simbel Temples	Temple	Aswan	140
Sacred Lake	Lake	Luxor	139
Mastaba	Tomb	Giza	136
Khufu Pyramid	Pyramid	Giza	124
Statue of Queen Hatshepsut	Statue	Luxor	124
Amenhotep III Template	Temple	Luxor	119
Queen Hatshepsut Mummy	Mummy	Cairo	106
Court of King Thutmose I	Heritage Site	Luxor	105
Bust Statue of Akhenaten	Statue	Luxor	75
The Great Sun Court of Aton	Heritage Site	Luxor	72
Head Statue of King Hatshepsut	Statue	Cairo	68
Tutankhamun Coffin-Tutankhamun Sarcophagus	Sarcophagus	Cairo	66
Statue of Tutankhamun with Ankhesenamun	Statue	Luxor	58
Statue of Akhenaten	Statue	Cairo	55
Tomb of King Den	Tomb	Sohag	48
King Thutmose II Mummy	Mummy	Cairo	43
Giza Pyramids	Pyramid	Giza	41
Goddess Isis with her child	Statue	Cairo	41
King Thutmose II	Statue	Luxor	21
Another Statue of Akhenaten	Statue	Cairo	20
Statue of Princess Meketaton	Statue	Cairo	16
Temple of Edfu	Temple	Aswan	11
41	11	6	7,778

that plateaued at 60%. The model's performance may benefit from more training epochs, as indicated by these results.

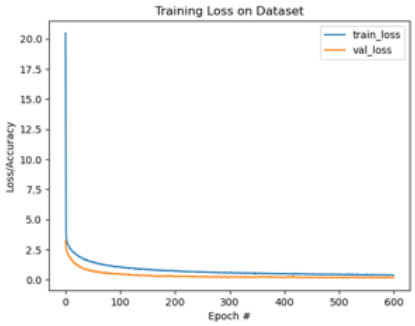
In Fig. 3(c), the graph demonstrates that the training and testing accuracy curves of the LeNet5 model display a consistent upward trend as the epochs' number increases. Notably, the testing accuracy is continuously more unstable than the training accuracy curve. Upon completion of the training process, it attains 94.29% for training accuracy and 92.64% for testing accuracy. Besides Fig. 3(f), when analyzing the

TABLE III. EVALUATION METRICS

Model	Type	Measurment			
		Accuracy	Precision	Recall	F1-Score
ResNet50	Train	**99.69%**	**99.73%**	**99.67%**	**99.70%**
InceptionV3		93.31%	96.52%	90.53%	93.37%
LeNet5		94.29%	95.20%	93.17%	94.16%
ResNet50	Test	**99.13%**	**99.28%**	**99.13%**	**99.20%**
InceptionV3		90.90%	87.90%	94.21%	90.87%
LeNet5		92.64%	94.13%	91.34%	92.69%



(a)



(b)

Fig. 4. (a) Accuracy of inceptionV3 model for 600 epochs (b) Loss curve of inceptionV3 model for 600 epochs.

loss graph, it is evident that, in the beginning, the loss during testing was lower than the loss during training. However, as the epochs' number progressed, both the training and testing loss curves showed a decreasing trend.

In terms of generalization, an experiment has been conducted with the ResNet50 model, as it is the highest accuracy among the three models. It has predicted new, unseen data of the same domain with an accuracy of 97.43%. Notedly, the unseen data size is over 35 thousand images suggesting that the dataset is scalable.

V. CONCLUSION

The Egypt Monuments Dataset v1 is introduced as a new benchmark for classification and instance recognition on a country-wide scale. Unlike many current computer vision datasets, this dataset has the following features: 1) it was gathered by individuals who are not computer vision professionals for a specific goal, making it unbiased; 2) unlike previous datasets, it is a better representation of real-world challenges; 3) it presents a classification challenge with a long-tail distribution; and 4) it has practical applications in the fields of conservation and Egyptology.

In terms of domain coverage, the EGYPT-V1 dataset demonstrates scalability by covering most of the famous pharaonic monuments allocated in Egypt. In addition, the proposed approach performs well across different subcategories. The approach's ability to perform well on a diverse range of data within the same domain suggests that the dataset is scalable.

Future plans involve undertaking object detection tasks. It is also planned to increase the size of the dataset where Egypt, in particular, is home to one-third of the world's monuments. This approach would promote the development of new methods for measuring errors. Finally, it is anticipated that this dataset will have value in examining how to train humans in recognizing intricate visual classes, and experimentation with human learning models is intended.

ACKNOWLEDGMENT

The authors extend their gratitude to Cercle, a live-stream media platform specializing in the film and broadcast of DJ sets and live performances within cultural heritage sites and landmarks through electronic music and video mediums. We would like to acknowledge Cercle for their collaboration in providing cultural heritage videos for our use. The authors also express their appreciation to WatchiT, a video-on-demand service for Arabic content online, with a unique collection of diverse entertainment options, including movies and TV shows, for their support of our work.

REFERENCES

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," *2007 IEEE conference on computer vision and pattern recognition*, pp. 1–8, 2007.

[2] H. Jgou, F. Perronnin, M. Douze, J. Snchez, P. Prez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[3] A. Gordo, J. Almazán, J. Revaud, and D. Larlus, "Deep image retrieval: Learning global representations for image search," *European conference on computer vision*, pp. 241–257, 2016.

[4] F. Radenović, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, "Revisiting oxford and paris: Large-scale image retrieval benchmarking," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5706–5715, 2018.

[5] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," *Proceedings of the 1st ACM international conference on multimedia retrieval*, pp. 1–7, 2011.

[6] R. Del Chiaro, A. D. Bagdanov, and A. Del Bimbo, "Noisyart: A dataset for webly-supervised artwork recognition." *VISIGRAPP (4: VISAPP)*, pp. 467–475, 2019.

[7] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," *Unpublished manuscript*, vol. 40, no. 7, pp. 1–9, 2010.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, pp. 211–252, 2015.

[9] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[10] J. Turner, *The Dictionary of Art*. Macmillan, 1996.

[11] A. Elnagar and A. Derbali, "The importance of tourism contributions in egyptian economy," *International Journal of Hospitality and Tourism Studies*, vol. 1, no. 1, pp. 45–52, 2020.

[12] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," *2008 IEEE conference on computer vision and pattern recognition*, pp. 1–8, 2008.

[13] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometry consistency for large scale image search-extended version," 2008.

[14] Y. Avrithis, G. Tolias, and Y. Kalantidis, "Feature map hashing: Sub-linear indexing of appearance and global geometry," *Proceedings of the 18th ACM international conference on Multimedia*, pp. 231–240, 2010.

[15] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," *European Conference on Computer Vision*, pp. 748–761, 2010.

[16] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[17] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys *et al.*, "City-scale landmark identification on mobile devices," *CVPR 2011*, pp. 737–744, 2011.

[18] Y. Li, N. Snavely, D. P. Huttenlocher, and P. Fua, "Worldwide pose estimation using 3d point clouds," *Large-Scale Visual Geo-Localization*, pp. 147–163, 2012.

[19] K.-H. Yap, Z. Li, D.-J. Zhang, and Z.-K. Ng, "Efficient mobile landmark recognition based on saliency-aware scalable vocabulary tree," *Proceedings of the 20th ACM international conference on Multimedia*, pp. 1001–1004, 2012.

[20] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, "24/7 place recognition by view synthesis," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1808–1817, 2015.

[21] T. Weyand and B. Leibe, "Visual landmark recognition from internet photo collections: A large-scale evaluation," *Computer Vision and Image Understanding*, vol. 135, pp. 1–15, 2015.

[22] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-scale image retrieval with attentive deep local features," *Proceedings of the IEEE international conference on computer vision*, pp. 3456–3465, 2017.

[23] V. Sharma, "Qutub complex monuments' images dataset," Oct 2018, Accessed: Sept. 23, 2022. [Online]. Available: https://www.kaggle.com/datasets/varunsharmaml/qutub-complex-monuments-images-dataset

[24] R. Gupta, P. Mukherjee, B. Lall, and V. Gupta, "Semantics preserving hierarchy based retrieval of indian heritage monuments," *Proceedings of the 2nd Workshop on Structuring and Understanding of Multimedia heritAge Contents*, pp. 5–13, 2020.

[25] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2575–2584, 2020.

[26] Y. Avrithis, Y. Kalantidis, G. Tolias, and E. Spyrou, "Retrieving landmark and non-landmark images from community photo collections," *Proceedings of the 18th ACM international conference on Multimedia*, pp. 153–162, 2010.

[27] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[28] S. Hesham, R. Khaled, D. Yasser, S. Refaat, N. Shorim, and F. H. Ismail, "Monuments recognition using deep learning vs machine learning," *2021 IEEE 11th annual computing and communication workshop and conference (CCWC)*, pp. 0258–0263, 2021.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.