Beyond Ensembles: Architecture-Level Fusion for Enhanced Monument Heritage Recognition

Mennat Allah Hassan[1]*, Mona M. Nasr[2], Alaa Mahmoud Hamdy[3]
Faculty of Computers and Artificial Intelligence, Helwan University, Egypt[1, 2]
Faculty of Computing and Information Sciences, Egypt University of Informatics, Egypt[1]
Faculty of Engineering, Helwan University, Egypt[3]

***Abstract*—Heritage is seen as a key part of nations, including a broad variety of traditions, cultures, monuments, plants and animals, foods, music, and further. Regarding countries, their own heritages are defined by preservation, excavation, and restoration of historical assets that are important and show the nation's history. It comprises a wide range of physical objects and materials found in cultural institutions which are moveable heritage, as well as the heritage found in built environments which are immovable and natural landscapes. Previous studies on monument classification frequently used single small datasets, limiting accuracy and generalizability. This work introduces a proposed model and a thorough experimental comparison of widely used deep learning architectures, specifically Convolutional Neural Networks and Transformers beside our proposed model, for monument recognition in the cultural monument domain. It seeks to conduct a comparative experiment for selecting representatives from these two methodologies regarding their capacity for transferring information from a general dataset, like ImageNet, to heritage landmarks datasets of varying sizes. When we tested samples of the topologies ResNet, DenseNet, and Swin Transformer (Swin-T), we find that the proposed model had the best results, however ResNet-50 achieved comparable accuracy to Swin-T.**

***Keywords*—Cultural monument; heritage landmarks; monument classification; monument recognition; transformers**

I. Introduction

Heritage is regarded as a fundamental aspect of a nation, encompassing a diverse range of traditions, cultures, monuments, flora and fauna, culinary practices, languages, music, and more [1]. For countries, their own heritages are characterized by the preservation, excavation, and restoration of historical artifacts that hold significant value and represent the legacy of the nation [2].

Cultural heritage is a multifaceted tapestry that reflects our past, present, and future aspirations [3]. It broadly includes a diverse array of tangible items and materials found in cultural institutions as movable heritage, as well as the heritage embodied in constructed environments (immovable) and natural landscapes. As noted in several discussions, cultural heritage promotion and preservation have enormous potential to make life better, boost the economy, and make societies that are lively, creative, and rich [4].

Machine Learning (ML) alongside Deep Learning (DL) are used for classifying monuments images. Image classification (IC) is crucial for the effective recognition and comprehension of the diverse range of monuments found worldwide. By leveraging information and communication technology (ICT) tools to develop a digital library of monuments, alongside artificial intelligence (AI) for identifying historical sites, we create a connection to the past, shedding light on the significance of their architecture and history. This research has important societal implications. Tourist experiences can be improved by better monument recognition, which makes cultural heritage more approachable for both visitors and residents. Additionally, this supports preservation efforts by facilitating the evaluation of structural and conservation requirements. Moreover, this approach aids urban planning, allowing planners to identify monument locations and integrate that information into city development strategies, assuring which modern buildings and infrastructure projects respect the city's heritage history [5–8].

Numerous studies have recently demonstrated effective image classification for a variety of applications. Despite the significance of this area, there is a notable lack of research focused specifically on the classification or recognition of tangible cultural heritage, such as monuments [9]. While DL has enabled automated monument recognition, existing approaches exhibit critical limitations: (1) reliance on single-architecture models that cannot capture complementary visual features, (2) evaluation on small datasets of <5,000 images limiting generalizability, and (3) lack of adaptive feature integration mechanisms [5-12].

In this work, we perform an experimental comparison of prevalent DL architectures, particularly Convolutional Neural Networks (CNNs) plus Transformers, for monument recognition within the cultural heritage domain. We use a fine-tuning strategy that is specific to the task of classifying monuments in our method. It is very important that the system can learn new tasks with only a few training samples, since getting a lot of annotated data is very expensive. We look at the architectures of ResNet-50 [10], DenseNet-121 [11], and Swin Transformer (Swin-T) [12] and compare their results to monument classification with respect to accuracy and computational complexity. We adjust and test these models on two datasets of heritage monuments.

1) Research problem: How can architecture-level fusion combine complementary CNN strengths for superior monument recognition while maintaining computational efficiency? CNNs excel at local spatial features while

*Corresponding author.

modernized ConvNets capture global context—monuments require both for distinguishing similar architectural structures.

2) Why fusion is necessary: EfficientNet-B4 uses compound scaling with squeeze-and-excitation attention for local textures and architectural details. ConvNeXt-Tiny employs depthwise separable convolutions for global semantic representations. Monuments require: (a) local features distinguishing similar styles, (b) global context for structural configurations, and (c) adaptive weighting based on input characteristics. No single architecture addresses all three requirements.

3) Contributions: This study offers the following contributions: *(1) Methodological:* We propose a dual-branch fusion architecture combining EfficientNet-B4 and ConvNeXt-Tiny with an adaptive feature fusion module. Unlike existing single-architecture approaches [9, 17, 20] or static ensemble methods [15, 24] that use fixed-weight averaging, our learnable gating mechanism dynamically adjusts branch contributions (α, β) based on input characteristics, enabling texture-focused or context-focused processing as needed.

4) Empirical: We conduct comprehensive evaluation across two heritage datasets of varying scales—Egypt-v1 (7,778 images, 41 classes) and Pisa (1,227 images, 12 classes)—and establish systematic baselines comparing ResNet-50, DenseNet-121, and Swin-T under consistent experimental protocols.

5) Practical: The proposed model achieves 99.77% accuracy on Egypt-v1, supporting real-world applications including automated monument cataloging for preservation agencies, enhanced mobile tourism guides, and urban planning systems that integrate heritage site data into city development strategies.

The rest of this paper is set up like this. In Section II, there is a comprehensive review of the literature. The tested architectures are in Section III. Then, section IV talks about the two datasets' features. While Section V and VI is all about the proposed model and experimental settings. Moreover, the results that were achieved are explained in Section VII. Section VIII provides discussion and interpretation of findings. Section IX concludes with limitations and future directions.

II. Comprehensive Literature Review

Research in DL-based monument recognition has evolved significantly over recent years, with approaches broadly categorized into three methodological paradigms: CNN-based methods, Transformer-based methods, and hybrid/ensemble approaches. This section systematically reviews representative works within each category, critically analyzes their limitations particularly regarding dataset scale, and identifies the research gaps that motivate our proposed fusion architecture, as shown in Table I and discussed.

Boyadzhiev et al. [13] performed a comparative analysis of several deep neural network architectures, which comprised both CNNs and Vision Transformers (ViTs), for the heritage image classification. Their results demonstrated the strong effectiveness of these models, reporting classification accuracies of approximately 96.8% for VGG11, 97.4% for ResNet34, 97.8% for DenseNet, 98.0% for PoolFormer, 97.9% for ViT, and 98.8% for Swin Transformer. While these findings offer insightful information about the relative performance of different architectures in cultural heritage applications, it had limitations due to the small dataset of only 1,227 images representing 12 monuments for the Pisa dataset [14].

Sasithradevi et al. [9] created the MonuNet model, which is a specialized deep learning model. MonuNet solves the problem of sorting through old pictures of Kolkata's important buildings. It was trained over a carefully chosen dataset of 13 heritage locations, each with 50 photos to make sure there was a fair representation. MonuNet used Dense and attention modules for parallel-spatial channels to make feature extraction and classification more accurate. The model did better than typical DenseNet models, with an accuracy of 89%, a precision of 86.77%, and a recall of 86.61%. These results demonstrated MonuNet's effectiveness in heritage image classification and its potential applications in cultural preservation, tourism, and urban planning. While MonuNet performed well in classifying Kolkata's heritage monuments, it had limitations due to the small dataset of only 50 images per site, which would affect generalizability.

Djelliout and Aliane [15] proposed a Multi-CNN model for the multi-classification of cultural historical monuments, addressing various dimensions such as monument identity, architectural type, and historical period. Using the AlgHeritage dataset [16], containing over 20k images of 90 distinct monuments, the Multi-CNN model integrated several CNN architectures including DenseNet169, MnasNet, and GoogleNet. The classification accuracy attained by the model was 94.52%, surpassing other single models like DenseNet169 with accuracy of 93.70%, MnasNet accuracy of 92.80%, and GoogleNet accuracy of 88.18%. These results indicated the superior performance of the Multi-CNN model in recognizing and categorizing heritage monuments, demonstrating its potential for applications in heritage conservation, documentation, and tourism. Despite its 94.52% accuracy, the Multi-CNN model's reliance on the AlgHeritage dataset limited generalization, and its computational complexity requires more resources compared to single models.

Khandelwal et al. [17] introduced a study focusing upon the effective classification of historical sites utilizing various CNN architectures. The authors tested ResNet50, InceptionResNetV2, EfficientNetB1, EfficientNetB3, and MobileNetV2 on a set of 24 Indian monuments. There were 4,895 photos in the dataset, and to make the model work better, data augmentation and hyperparameter tuning were applied. MobileNetV2 was the best of the models examined, with 95.58% for the validation accuracy and 99.90% for the training accuracy. It showed which is the best model for classifying monuments. Their work demonstrated the transfer learning and fine-tuning work well regarding the monument recognition. This means that deep learning models like MobileNetV2 can classify objects with a high degree of accuracy with only a few parameters, making them useful for real-time applications. MobileNetV2 had a validation accuracy of 95.58%, which was better than other models. However, it relied on data augmentation and hyperparameter tuning, which shows that it

is not very scalable. The model also showed evidence of mild overfitting because it was trained on a dataset that was not very big.

Kukreja et al. [18] proposed a hybrid DL model regarding the multi-classification of Indian cultural sites utilizing a real-phase image dataset. The model combined CNNs alongside with Long Short-Term Memory (LSTM) networks for classifying images of monuments. They did two key things: they did a binary classification of heritage along with non-heritage monuments, which was 92.37% accurate, and a multi-classification task that divided monuments into four groups, which was 95.89% accurate. This hybrid model showed high efficiency in monument recognition as well as classification, supporting the preservation and awareness of cultural heritage. Kukreja et al. used a dataset of 3,000 images from four prominent Indian monuments, and their model outperformed traditional classification methods in respect to accuracy as well as classification performance, as demonstrated by precision, recall, and F1 scores. Despite achieving 95.89% accuracy in multi-classification, their proposed hybrid model faced limitations due to the small dataset of only 3,000 images, which would hinder its ability to generalize to a broader range of monuments.

TABLE I. LITERATURE REVIEW

Ref.	Dataset	Model	Metric	Result in Percentage format	Limitation
[13]	Pisa dataset [14]	VGG, ResNet, DenseNet, PoolFormer, ViT, Swin-T	Accuracy	VGG11≈ 96.8, ResNet34 ≈ 97.4, DenseNet ≈ 97.8, PoolFormer ≈ 98.0, ViT ≈ 97.9, **Swin-T ≈ 98.8**	A limitation is that its small dataset of only 1,227 images representing 12 monuments, may limit its generalization to other monuments or larger datasets.
[9]	Heritage Sites in Kolkata	MonuNet, DenseNet201, DenseNet169, DenseNet121	Accuracy, Precision, Recall	**MonuNet, 89, 87.8, 86.6** DenseNet201, 79, 73.5, 74.6 DenseNet169, 85, 83.2, 82.6 DenseNet121, 85, 83.2, 81.6	Limited by the small dataset of only 50 images across 13 sites, which may affect generalizability.
[15]	AlgHeritage dataset [16]	GoogleNet, Densenet169, MnasNet, Multi-CNN	Accuracy	GoogleNet, 73.1 Densenet169, 88.2 MnasNet, 86.7 **Multi-CNN, 92.1**	The proposed model's reliance on the dataset limits generalization, and its computational complexity requires more resources compared to single models.
[17]	24 types of Indian monuments	ResNet-50, InceptionResNetV2, EfficientNetB3, EfficientNetB1, MobileNetV2	Accuracy	ResNet-50, 12.8 InceptionResNetV2, 99.7 EfficientNetB3, 91.8 EfficientNetB1, 95.4 **MobileNetV2, 99.9**	The model showed signs of moderate overfitting due to the relatively small dataset used for training.
[18]	Indian heritage monument dataset	CNNs With LSTM	Precision, Recall, F1-score	Class One, 92.6, 95.2, 93.8 Class Two, 95.1, 94.1, 94.6 Class Three, 95.4, 95, 93.2 Class Four, 96.1, 93.4, 93.9	Limited by the small dataset of only 3k images, which may hinder its ability to generalize to a broader range of monuments.
[19]	Real-phase Indian dataset	MLP	Precision, Recall, F1-score	Class One, 95.8, 98.7, 96.4 Class Two, 95.6, 89.3, 91.9 Class Three, 91.9, 93.6, 92.0 Class Four, 95.6, 93.7, 94.0	Limited by the small dataset of 10k images, which may restrict its ability to generalize across more diverse heritage categories.
[20]	Egypt Monuments Dataset v1	ResNet50, Inception V3, LeNet5	Accuracy	**ResNet-50, 99.1** InceptionV3, 90.9 LeNet5, 92.6	It is relatively small size of 7,778 images, which may hinder the generalization of models when applied to a broader set of monuments beyond the Egyptian context.
[21]	UMS landmark dataset [22], Scene-15 dataset [23]	EFFNET 1, EFFNET 2, RESNet152	Accuracy	EFFNET 1: {LSVM:100, CNN (2D): 100, CNN (1D): 100, GBDT: 100,SGD: 100, MLP: 44} EFFNET 2: {LSVM: 94, CNN (2D), 95, CNN (1D): 100, GBDT: 100, SGD: 100 , MLP: 12} RESNet152: {LSVM: 100, CNN (2D): 85, CNN (1D): 100, GBDT: 100} EFFNET 1: {LSVM: 94, CNN (2D): 85, CNN (1D): 94, GBDT: 68, SGD: 68, MLP: 43} EFFNET 2: {LSVM: 94, CNN (2D): 91, CNN (1D): 92, GBDT: 66, SGD: 92, MLP: 40} RESNet152: {LSVM: 62, CNN (2D): 58, CNN (1D): 62, GBDT: 41}	A limitation of the proposed model is that while it achieves high accuracy, the extra pre-processing for feature reduction increases computational overhead, which may affect scalability in larger or more complex environments.
[24]	Indian dataset of 4.5k heritage palaces	Hybrid CNN-SVM model	Precision, Recall, F1-score	Class One, 88.7, 69.4, 93.8 Class Two, 84.9, 71, 94.6 Class Three, 85.2, 72.4, 93.2 Class Four, 81.1, 72.4, 93.9	A limitation is that its small dataset of 4.5k images may limit its generalization to other monuments or larger datasets.

Kukreja et al. [19] employed a DL-based Multi-Layer Perceptron (MLP) model regarding the multi-classification of heritage Indian images. It was trained over a dataset of 10k images, categorized into four heritage classes: animals, birds, monuments, and paintings. Their work utilized data augmentation methods to improve the dataset, and the MLP model achieved notable results. In the binary classification task, differentiating among the heritage and non-heritage images, their model attained an accuracy of 94.32%. For the multi-classification task, the model achieved an accuracy of 95.43%, with animal heritage images yielding the highest performance metrics, including a precision of 95.84%, recall of 98.65%, and F1-score of 96.37%. These outputs showed that the model worked well for recognizing and classifying heritage, which helps protect and raise awareness of cultural heritage using digital solutions. Despite achieving 95.43% accuracy in multi-classification, the proposed MLP model proposed was limited by the small dataset of 10k images, which would restrict its ability to generalize across more diverse heritage categories.

Hassan et al. [20] introduced the Egypt Monuments Dataset v1, a comprehensive for image classification plus instance-level recognition of Egyptian monuments and heritage sites. This dataset consists of 7,778 images across 41 categories, including famous monuments for example tombs, heritage-sites, and statues. The authors evaluated the performance of several DL models, including ResNet50, Inception V3, and LeNet5, on this dataset. ResNet50 achieved the highest accuracy with 99.13%, followed by LeNet5 with 92.64%, and Inception V3 with 90.90%. These results demonstrated the dataset's potential for advancing the classification and recognition of heritage monuments, particularly with respect to real-world applications in Egyptology and cultural heritage preservation. Despite the strong performance of ResNet50, achieving 99.13% accuracy, the Egypt Monuments Dataset v1 faced limitations due to its relatively small size of about 7k images, which would hinder the generalization of models when applied to a broader set of monuments beyond the Egyptian context.

Razali et al. [21] developed a lightweight DL-based landmark recognition model for smart tourism integrating CNN with Linear Discriminant Analysis (LDA). It was trained over the UMS Landmark dataset [22] and the Scene-15 dataset [23] to identify tourist landmarks and public scenes. The best feature extractor was EfficientNet (EFFNET), which got a flawless classification accuracy of 100% on the UMS dataset and 94.26% on the Scene-15 dataset. Additionally, the use of LDA decreased the number of dimensions of the features by over 90% avoiding compromising classification performance. This approach demonstrated a significant reduction in computational complexity while preserving a high level of accuracy, which makes it perfect for smart tourism applications in real time. A limitation of their proposed model was that while it achieved high accuracy, the extra pre-processing for feature reduction increased computational overhead, which would affect scalability in larger or more complex environments.

Kumar et al. [24] employed a hybrid approach integrating CNN with Support Vector Machines (SVM) for the multi-classification of Indian heritage palaces. This dataset, consisting of 4,500 images of various heritage palaces, was pre-processed and divided to 75% training and 25% testing sets. This hybrid CNN-SVM model achieved impressive results, with a classification accuracy of 97% for features like grand fountains and Doric pillars. Precision, recall, and F1-scores were also evaluated, with Class 1 which is grand fountains achieving a precision of 88.73% and recall of 69.44%, while Class 2 which is Doric pillars showed a precision of 84.94% and recall of 71.01%. This work demonstrated the effectiveness of the hybrid model for heritage monument classification and offered a valuable tool for cultural heritage preservation and analysis. The limitation of Kumar et al.'s model was that its small dataset of 4.5k images would limit its generalization to other monuments or larger datasets.

The literature reveals four critical gaps motivating our fusion approach:

1) Feature complementarity neglect: Existing studies evaluate single architectures in isolation. CNNs excel at local textures and spatial hierarchies, while modernized ConvNets and Transformers capture global relationships. Monument recognition requires both capabilities, yet no work combines architectures with complementary inductive biases.

2) Static feature integration: Hybrid approaches use naive strategies (ensemble voting, concatenation, fixed-weight averaging) that cannot adapt to input characteristics. Effective fusion requires adaptive mechanisms that dynamically weight contributions based on whether an image demands textural detail or global context understanding.

3) Insufficient dataset scale analysis: Studies report results on single datasets without examining how architectures scale across varying data availability (650 to 20,000+ images), conflating memorization with generalization.

4) Lack of modern architecture combinations: State-of-the-art architectures (EfficientNet's compound scaling, ConvNeXt's modernized design) remain unexplored in fusion configurations despite offering complementary strengths suited for monument recognition.

In summary, CNNs sacrifice global context for local discrimination, Transformers exhibit the inverse trade-off, and existing hybrids lack adaptive integration. These gaps motivate our dual-branch fusion combining EfficientNet-B4 and ConvNeXt-Tiny with learnable gating and channel attention.

III. Tested Architectures

CNNs and Transformers architectures are the two primary deep learning paradigms that have dominated computer vision research in recent years. CNNs are distinguished by their strong generalization ability in image-related tasks and comparatively low computational cost. Convolutional layers' translation-invariance and locality characteristics, which offer a potent inductive bias, are the source of this effectiveness. Transformer-based models have attracted much interest lately for their capacity to use attention mechanisms for capturing global relationships and long-range dependencies. However, attention layers' scalability in practical applications are limited by their computational complexity, which increases quadratically with input size. In order to solve this, Tay et al. [25] have proposed a number of effective Transformer variants that maintain competitive performance by substituting lighter

alternatives. Despite these advances, Transformers are often less robust in terms of generalization and typically require large-scale pre-training to achieve strong results Csordás et al. [26]. Fortunately, pre-trained models from both paradigms are widely available and can be fine-tuned on domain-specific datasets, enabling their application to diverse real-world scenarios.

In this study, to highlight the task of monument recognition for cultural heritage applications, we go beyond the limitation of previous studies that relied solely on the Pisa dataset. To ensure greater diversity and robustness, we incorporate two additional monument datasets alongside Pisa, enabling a broader and more challenging benchmark. We focus our comparative experimental analysis on three representative architectures: ResNet-50, DenseNet-121, and Swin-T. Each model is optimized for the particular monument classification tasks after being pre-trained on ImageNet, allowing us to compare their effectiveness and transferability across different datasets. The ability to adapt to new tasks from few available training samples remains crucial, given the cost and difficulty of collecting large-scale annotated cultural heritage datasets. Furthermore, Table II shows how many parameters and how much computing power in Floating Point Operations per Second (FLOPS) each model needs to process data. Researchers typically use FLOPS to compare models. A model with lower FLOPS is lighter, faster, and more efficient, whereas a model with higher FLOPS is heavier, slower, and uses more resources [10-12].

TABLE II. COMPARISON OF THE NUMBER OF PARAMETERS AND THE COMPUTING COST (IN FLOPS) OF DIFFERENT MODELS

Model	***# Parameters***	***FLOPS***
ResNet-50 (v1)	~ 25M	3.80×10^9
DenseNet-121	~ 8M	2.91×10^9
Swin-T	~ 28M	4.50×10^9

IV. DATASETS CHARACTERISTICS

Two datasets of cultural monuments or heritage landmarks were considered for this study: the Egypt Monuments Dataset v1 (EGYPTv1) [20] and the Pisa Dataset [14]. *Dataset Selection Rationale:* These datasets were strategically selected based on four criteria: (1) Scale diversity—Egypt-v1 (7,778 images, 41 classes) represents a large-scale dataset while Pisa (1,227 images, 12 classes) represents a small-scale dataset, enabling evaluation of model behavior across varying data availability conditions; (2) Geographic and cultural diversity—the datasets cover distinct heritage contexts (Ancient Egyptian monuments vs. Italian Renaissance/Medieval architecture), testing cross-cultural generalization; (3) Public availability and reproducibility—both datasets are publicly accessible, enabling reproducible research; (4) Benchmark relevance—Egypt-v1 is the first dedicated Egyptian heritage dataset with established baselines [20], while Pisa is a widely-used benchmark in cultural heritage recognition literature [13, 14]. This dual-dataset approach addresses a key limitation of prior studies that evaluated on single datasets, conflating memorization with generalization. More details will be provided in the next subsection.

A. EGYPT-v1 Dataset

A benchmark standard for ILR and fine-grained IC across the ancient Egyptian monuments field, is the Egypt-v1 dataset, which was first presented by Hassan et al. [20]. It is the first dataset devoted to Egyptian heritage sites, containing 7,778 photos from 6 of Egypt's 28 governorates, representing 41 different monument classes. Luxor is home to about 37% of the monuments with more than 2,000 photos, whereas Cairo is home to about 27%. The dataset includes various categories such as pyramids, temples, statues, busts, and heritage sites, sourced manual and semi-automated primarily from different platforms like YouTube, and Wikimedia Commons. The images reflect diverse conditions, including indoor and outdoor settings, various lighting scenarios, and angles, enhancing the applicability of the dataset to real-world challenges, as shown in Fig. 1.

Three deep learning models, ResNet50, InceptionV3, and LeNet5, were evaluated in EGYPT-v1, achieving test precisions of 99.13%, 90.90%, and 92.64%, respectively. ResNet50 demonstrated the highest performance and scalability, achieving 97.43% accuracy on unseen data with over 35,000 images. This dataset supports a broad range of applications, including conservation and Egyptology, and provides a foundation for future work in monument recognition, such as object detection and larger-scale expansions.

B. Pisa Dataset

The Pisa Dataset, introduced by Amato et al. [14], is a curated collection of 1,227 images depicting 12 cultural heritage sites and monuments in Pisa, Italy. These images were sourced from the online photo-sharing platform Flickr, and their corresponding IDs and labels are publicly accessible at https://falchi.isti.cnr.it/pisaDataset/, as illustrated in Fig. 2.

The dataset was developed to support research on monument recognition in images, a task that presents challenges due to variations in viewpoint, lighting, and image quality. To address this, the authors explored k-Nearest Neighbors (kNN) based classification plus landmark recognition techniques, proposing two novel methods that combine kNN with local visual descriptors. Notably, through the use of similarity search access methods, their first approach makes it possible to perform standard kNN classification more efficiently by introducing a more lenient definition of image-to-image similarity based on local features.

Fig. 1. Samples from three classes of the EGYPTv1 benchmark dataset [20].

Fig. 2. Samples from 12 classes of the Pisa dataset [14].

V. Proposed Model

Monument recognition is a difficult challenge for computer vision since models have to identify the difference between architectural objects that look identical while dealing with changes in viewing angles, illumination, occlusion, and scale. Traditional single-architecture methods, whether they use pure CNNs or Vision Transformers, frequently have trouble getting all the visual features needed for strong monument categorization. CNNs are great at finding local spatial hierarchies and textural patterns because of their inductive biases. Modern ConvNet [28] systems, like ConvNeXt, use design ideas from Vision Transformers to get more global contextual information. The proposed work presents an innovative dual-branch fusion design that synergistically integrates EfficientNet-B4 [29] and ConvNeXt-Tiny, capitalizing on the complementing capabilities of both architectural paradigm as shown in Fig. 3. EfficientNet-B4 uses efficient compound scaling and squeeze-and-excitation attention mechanisms to learn discriminative local features well. ConvNeXt-Tiny, on the other hand, uses modernized convolutional designs with depthwise separable convolutions and layer normalization to make stronger semantic representations. This paper suggests an adaptive feature fusion module with learnable gating mechanisms and channel attention that dynamically weights the contribution of each branch based on the characteristics of the input, rather than just combining features from both networks. This smart fusion method lets the model concentrate EfficientNet features for areas with a lot of texture and ConvNeXt features for scenes that need a richer awareness of the context. This leads to better monument recognition over a wide range of datasets.

A. Dual-Branch Features

The proposed architecture uses a parallel dual-branch topology in which both branches process the same input image at the same time through separate pathways. The first branch uses EfficientNet-B4, a compound-scaled convolutional neural network that uses a principled compound coefficient to systematically change the depth, breadth, and resolution of the network. EfficientNet-B4 uses mobile inverted bottleneck convolution (MBConv) blocks with squeeze-and-excitation (SE) modules that change the responses of features in each channel by using global average pooling and then two fully connected layers with sigmoid activation. This architecture takes 224×224 RGB images and processes them via seven stages of MBConv blocks, each of which reduces the image's spatial resolution. In the end, it creates 1792-dimensional feature representations that contain detailed information on the image's spatial hierarchies and textures.

The second branch includes ConvNeXt-Tiny, which is a modernized version of standard ConvNet architecture that uses certain design ideas from Vision Transformers while still being able to conduct convolutions efficiently. ConvNeXt uses depth-wise separable convolutions with bigger 7×7 kernels, inverted bottleneck structures where channel expansion happens in the

middle layers, and GELU activation functions instead of ReLU. ConvNeXt makes a "purer" convolutional route by replacing batch normalization with layer normalization and using fewer activation functions and normalizing layers. The ConvNeXt-Tiny version processes the same 224×224 input through four steps of hierarchical feature extraction. This creates 768-dimensional feature vectors that capture more global semantic information and contextual links than typical CNNs. Both branches start with ImageNet pre-trained weights, which provide them robust feature representations learnt from more than 1.2 million pictures in 1000 categories. This transfer learning method speeds up convergence and makes generalization better, which is especially useful for monument recognition because training datasets are usually smaller than those for generic object recognition. The dual-branch design makes it possible to do both branches at the same time, which makes it possible to use the GPU efficiently through batch processing while keeping the computation manageable. The total inference time is still similar to that of single-architecture techniques because the dual-branch design may be done in parallel.

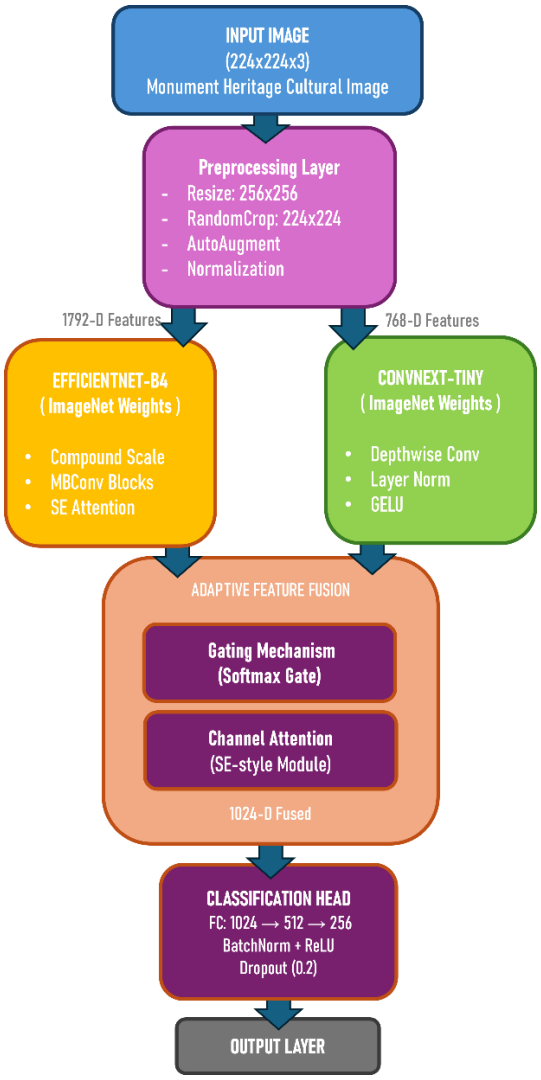

Fig. 3. Our proposed model architecture.

B. Adaptive Feature Fusion Mechanism

The main novel aspect about this architecture is the adaptive feature fusion module. This module uses a smart gating and attention mechanism to merge the different feature representations from EfficientNet-B4 of 1792-D and ConvNeXt-Tiny of 768-D. This module learns to change how much each branch contributes based on the input characteristics, which is different from naive concatenation or simple weighted averaging. This lets the model focus on EfficientNet features for texture-heavy monument details or ConvNeXt features when a broader architectural context is more useful.

The fusion process starts with two separate projection layers that turn both feature vectors into a shared 1024-dimensional space. This is done with fully connected layers, batch normalization, ReLU activation, and dropout regularization (rate=0.3). These projections make sure that the dimensions are compatible and let each branch learn how to change features in a way that is specific to the task. At the same time, the initial heterogeneous features which are 1792-D and 768-D, are combined to make a 2560-dimensional vector that goes into a gating network.

The gating network has a bottleneck architecture with two fully connected layers. The first layer uses ReLU activation and light dropout which is 0.15 to compress the 2560-D concatenated features to 640 dimensions (4× reduction). The second layer projects to 2 dimensions that show how important each branch is. A softmax activation makes sure that the values of these gates add up to one. This makes a normalized weighting scheme where α is the EfficientNet contribution and β is the ConvNeXt contribution where $\alpha + \beta = 1.0$. These learnt gate values change the projected features by multiplying them by each other, creating weighted representations that adaptively highlight the most informative branch for each input sample.

After gating, the weighted features from both branches are combined to make a 2048-dimensional representation. This representation then goes through channel-wise attention refinement, which is based on Squeeze-and-Excitation networks. This attention module has a global average pooling operation, followed by two fully connected layers. The first layer has an 8× channel reduction ratio, while the second layer has ReLU activation. The last layer has sigmoid activation for the final attention weights. These learned attention weights do channel-wise recalibration by lowering the importance of less informative feature channels and raising the importance of more discriminative ones. Finally, a projection layer uses a fully connected layer with batch normalization, ReLU activation, and dropout of 0.3 to turn the attention-modulated 2048-D features into a smaller 1024-D fused representation. This is the final unified feature vector that goes into the classification head.

This multi-stage fusion strategy—projection, gating, concatenation, attention, and final projection—lets the model do advanced feature integration that goes much beyond basic combination methods. The learnable gating mechanism makes the model easier to understand by showing which architectural branch has a bigger effect on predictions for different forms of

input. The channel attention, on the other hand, allows for more precise feature refining within the fused representation.

C. Classification Head

The classification head uses a three-layer multi-layer perceptron (MLP) with successive dimensionality reduction to turn the 1024-dimensional fused features into class predictions. The first fully connected layer goes from 1024 to 512 dimensions. Then, batch normalization is used to make training more stable, ReLU activation is used to make it non-linear, and dropout regularization where the rate equals 0.2, is used to stop overfitting. The second layer also normalizes, activates, and regularizes the data in the same way, but it cuts down on the number of dimensions from 512 to 256. The last layer makes logits for the 10 monument classes by projecting from 256 dimensions in a straight line.

Xavier (Glorot) uniform initialization is used to set the weights for all fully connected layers in the classification head. This method uses a uniform distribution with variance that is scaled by the number of input and output units. This way of starting helps keep the activation magnitudes the same across layers and speeds up convergence. The bias terms start off at zero. The three-layer design is not too deep, so it balances expressiveness with regularization. This avoids the problems of diminishing returns and overfitting that come with deeper classification heads while still giving it enough capacity to learn complex decision boundaries in the 1024-dimensional fused feature space.

Dropout with a modest 0.2 probability during training enables unpredictable regularization by randomly zeroing activations. This forces the network to learn redundant representations and makes it better at generalizing. Batch normalization layers use running statistics calculated across mini-batches to normalize activations. This lowers internal covariate shift and makes it possible to use larger learning rates. Although there are some theoretical concerns about how batch normalization and dropout work together, they do improve performance in this architecture by providing complementary regularization effects. Batch normalization deals with distribution shift while dropout encourages feature redundancy.

D. Training Methodology and Optimization

The proposed model uses an end-to-end fine-tuning technique, which means that all layers, including the pre-trained EfficientNet-B4 and ConvNeXt-Tiny backbones, can be trained from the beginning of training. This method is different from frozen-backbone or gradual unfreezing methods since it lets the whole network change its representations to better recognize monuments. End-to-end fine-tuning usually works better because it lets low-level feature extractors specialize for the target domain. However, it needs greater computational power and careful regularization to avoid catastrophic forgetting of pre-trained features.

The ImageNet pre-trained weights give a good starting point by encoding general visual ideas gained from real photos, such as edges, textures, and sections of objects. Monument images have a lot in common with ImageNet when it comes to how they look, such as building textures, geometric patterns, and outside scene aspects. This makes transfer learning work very well. However, monument-specific elements such as old stone weathering patterns, hieroglyphic details, and particular architectural styles necessitate domain adaptation, hence validating the comprehensive fine-tuning approach.

The training procedure uses many regularization methods to find a compromise between adaptability and keeping pre-trained knowledge. These include dropout in the fusion module of 0.3 and classifier of 0.2, label smoothing in the loss function which is ε=0.1, and weight decay in the optimizer of 0.01. These strategies stop overfitting while letting the model get better at recognizing monuments. The dropout rates are not sufficient, which keeps the quality of the pre-trained features while still giving the benefits of regularization.

With a smoothing parameter of ε=0.1, label smoothing cross-entropy loss trains the model. This stops the network from being too sure of its predictions. Standard cross-entropy with hard labels makes the model want to give the right class a probability of 1.0 and all other classes a probability of 0.0. This could lead to predictions that are overly confident, which hurts generalization and calibration. Label smoothing changes hard targets [0, 0, 1, 0, 0, ...] into soft targets. The true class gets a probability of 1-ε, whereas the wrong classes each get ε/(K-1), where K is the number of classes.

For monument recognition with 10 classes and ε=0.1, the true class gets a probability of 0.90, and each of the 9 wrong classes gets a probability of about 0.011. This smoothing has a number of benefits: it stops the model from pushing logits to extreme values, which makes the predicted probabilities more accurate; it encourages the representations in the penultimate layer to group more closely around class centroids; and it acts as a kind of implicit regularization that helps the model work better on unseen data.

VI. Experimental Setting

All the experiments are implemented with the two architecture-for monument recognition tasks using stratified k-fold cross-validation on two heritage cultural datasets with comprehensive GPU-optimized training infrastructure. These experiments are performed using PyTorch v2.7.1+cu118, Python v3.9, GPU of NVIDIA GeForce RTX 3060 Ti and TorchVision v0.22.1+cu118 and augmentation is done using TorchVision v2 API. Each network architecture was pre-trained on ImageNet, along with it was fine-tuned over two datasets.

1) Data preparation: It involves cleaning images with minimum 200-pixel shorter edge requirements, stratified splitting into 80% train-test and 20% validation sets, with further 80-20 subdivision of train-test data for each fold, ensuring balanced class distributions across all partitions.

2) Baseline: It utilizes ImageNet pre-trained ResNet50, DenseNet121 architectures alongside the final fully-connected layer replaced and Xavier-initialized for the specific number of monument classes identified in the two datasets respectively. Fine-tuning employs end-to-end training of the entire network with transfer learning from ImageNet weights, allowing all layers to be updated during training rather than freezing backbone features.

3) Data augmentation: It applies comprehensive transformation pipeline including resize to 256x256, random resized crop to 224x224 with scale range 0.533-1.875, AutoAugment [27] with ImageNet policy, random horizontal flipping, color jitter (brightness= 0.1, contrast= 0.1, saturation= 0.1, hue= 0.05). Finally, each image is normalized using ImageNet statistics where the mean equals [0.485, 0.456, 0.406]; while the standard deviation equals [0.229, 0.224, 0.225]).

4) Training algorithm: It uses the AdamW optimizer with $1e^{-4}$ for a learning rate, $1e^{-2}$for a weight decay, 32 for a batch size, cosine annealing for learning rate scheduling, label smoothing cross-entropy loss of ε=0.1, and mixed precision training with gradient scaling. It runs for 10 demo epochs per fold with a GPU-resident dataset architecture that pre-loads all images directly to GPU memory to avoid data loading bottlenecks and get the best computational efficiency.

VII. Results

The experimental assessment encompassed three deep learning baselines—ResNet-50, DenseNet-121, and Swin-T—and our suggested fusion model, characterized as a dual-branch fusion architecture that synergistically combines EfficientNet-B4 and ConvNeXt-Tiny. We tested these models on two datasets: Egypt-v1, which has 41 classes and more than 7,000 photos, and Pisa, which has 1,227 images of 12 monuments. We used standard classification metrics to quantify performance on both the testing and validation sets.

1) Overall performance: The proposed dual-branch fusion model achieved the highest testing accuracy on Egypt-v1 at 99.77%, marginally outperforming ResNet-50 of 99.75% and Swin-T of 99.59%. The proposed model demonstrated exceptional balance with precision of 99.69%, recall of 99.21%, and F1-score of 99.35%. On the Pisa dataset, Swin-T achieved the best testing performance with 99.22% accuracy, followed by DenseNet-121 of 97.66%, ResNet-50 of 96.88%, and the proposed model of 96.43%. However, the proposed model excelled in validation on Pisa with 98.93% accuracy, surpassing Swin-T of 98.75% and ResNet-50 of 98.12%, indicating superior generalization capability on this smaller dataset, as shown in Fig. 4.

2) Model comparison: As explained in Table III, DenseNet-121 showed the weakest performance on Egypt-v1 with only 92.19% testing accuracy and significant precision challenges of 84.42%, resulting in the lowest F1-score of 85.86%. However, it demonstrated a notable recovery on Pisa, achieving 97.66% testing accuracy and improving to 96.25% validation accuracy. Swin-T exhibited strong and consistent performance across both datasets, achieving near-identical validation accuracy to ResNet-50 on Egypt-v1 both at 99.67% while demonstrating exceptional metrics on Pisa with balanced precision-recall trade-offs precision of 99.56%, recall of 99.36%. ResNet-50 maintained robust performance with 99.75% testing accuracy on Egypt-v1 and strong generalization evidenced by its 99.67% validation accuracy.

TABLE III. Comparison of Model Performance Metrics, Underlined Text Shows the Best Values

Model	Dataset	Testing				Validation			
		Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
ResNet-50	Egypt-v1	99.75	99.71	99.83	99.76	99.67	99.39	99.51	99.44
	Pisa	96.88	97.61	94.50	95.55	98.12	98.12	95.91	96.53
DenseNet-121	Egypt-v1	92.19	84.42	90.91	85.86	93.75	95.04	93.18	91.14
	Pisa	97.66	98.37	95.19	96.02	96.25	97.08	93.55	94.74
Swin-T	Egypt-v1	99.59	99.40	98.48	98.71	99.67	99.51	99.67	99.56
	Pisa	99.22	99.56	99.36	99.44	98.75	99.17	96.67	97.48
Our Proposed	Egypt-v1	99.77	99.69	99.21	99.35	99.61	99.55	99.12	99.32
	Pisa	96.43	95.88	95.04	95.27	98.93	98.83	98.97	98.84

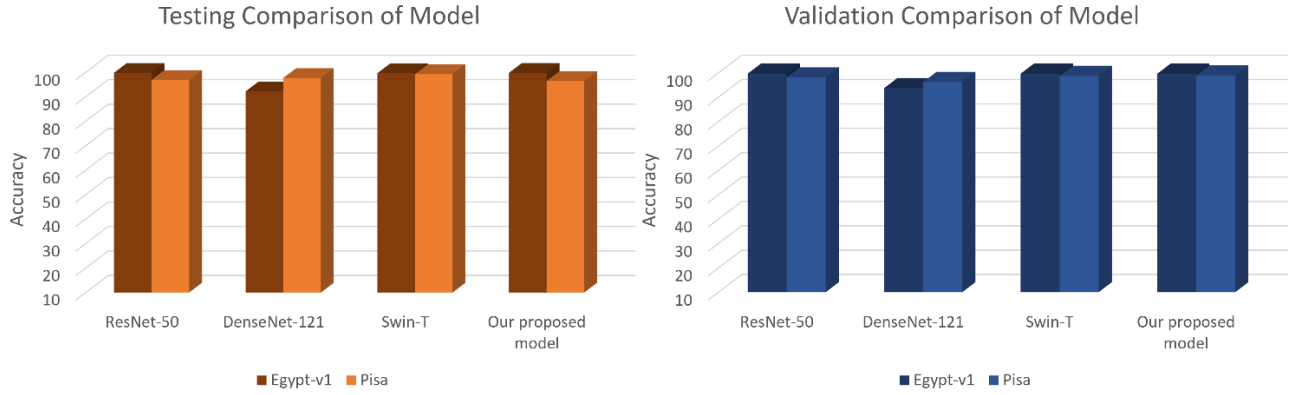

Fig. 4. Accuracy evaluated of each fine-tuned network using the testing and validation data.

3) Performance consistency: Most models maintained strong balance between precision and recall, with F1-scores closely aligned to overall accuracy metrics. The validation results demonstrated consistent performance patterns, with the proposed model showing particularly strong generalization on Pisa where the validation accuracy is 98.93% and the testing is 96.43%, indicating effective handling of the smaller dataset. All top-performing models which are ResNet-50, Swin-T, and the proposed model achieved validation accuracies above 99.6% on Egypt-v1, confirming robust generalization on the larger, more diverse dataset. Notably, the proposed model achieved the best testing-validation consistency on Egypt-v1 with only a 0.16% gap, demonstrating minimal overfitting.

VIII. Discussion

1) Interpretation of results: The proposed dual-branch fusion model achieved the highest accuracy (99.77%) on Egypt-v1, demonstrating that adaptive architecture-level fusion effectively captures complementary features—local textures via EfficientNet-B4 and global context via ConvNeXt-Tiny. The minimal testing-validation gap (0.16%) indicates the learnable gating mechanism successfully prevents overfitting by dynamically balancing branch contributions rather than relying on fixed weights.

2) Performance variation across datasets: On the smaller Pisa dataset, Swin-T outperformed our model in testing accuracy (99.22% vs. 96.43%), suggesting that transformer architectures may be more effective when training data is limited due to their pre-trained global attention mechanisms. However, our model's superior validation accuracy (98.93% vs. 98.75%) indicates better generalization capability. This discrepancy suggests the fusion approach benefits more from larger, diverse datasets where complementary feature learning can be fully exploited, while transformers leverage their extensive pre-training when fine-tuning data is scarce.

3) Comparison with existing methods: Unlike static ensemble approaches [15, 24] that assign fixed weights to each architecture, our adaptive gating mechanism learns input-dependent weights (α, β). This explains the consistent performance across diverse monument types—textured monuments (carved temples, weathered stone surfaces) benefit from higher EfficientNet contribution, while structurally distinct monuments (pyramids, towers) leverage ConvNeXt's global context understanding. DenseNet-121's poor performance on Egypt-v1 (92.19%) but recovery on Pisa (97.66%) suggests its dense connectivity pattern is better suited for smaller datasets with fewer classes.

4) Computational trade-offs: The ~60M parameter count represents a practical limitation for edge deployment compared to ResNet-50's 25M parameters. However, the parallel dual-branch design enables efficient GPU utilization, maintaining inference times comparable to single-architecture approaches. For resource-constrained applications such as mobile tourism guides, knowledge distillation could compress the model while preserving fusion benefits.

5) Implications for heritage preservation: The high accuracy achieved on Egypt-v1 supports practical deployment in automated monument cataloging systems, reducing manual annotation effort for preservation agencies. The model's robustness across varying lighting conditions and viewpoints (evidenced by Egypt-v1's diverse image sources from YouTube and Wikimedia Commons) suggests suitability for real-world tourism and documentation applications where image quality varies significantly.

6) Limitations of the current study: Several factors may limit the generalizability of our findings: (1) both datasets represent well-documented heritage sites with relatively clear images—performance on degraded or partially occluded monuments remains unexplored; (2) the evaluation is limited to classification tasks without addressing detection or segmentation; (3) the two datasets, while geographically diverse, represent Western and Middle Eastern heritage traditions—Asian, African, and American heritage sites may exhibit different visual characteristics requiring further validation.

IX. Conclusion and Future Work

This study addressed the limitations of single-architecture approaches for monument recognition by proposing a dual-branch fusion architecture combining EfficientNet-B4 and ConvNeXt-Tiny with adaptive feature integration. The proposed approach provides a foundation for scalable heritage recognition systems that can support global cultural preservation efforts, enhance tourist experiences through accurate mobile guides, and assist urban planners in integrating heritage considerations into city development.

1) Key findings: The proposed model achieved 99.77% accuracy on Egypt-v1, outperforming ResNet-50 (99.75%) and Swin-T (99.59%), with the best testing-validation consistency (0.16% gap), demonstrating minimal overfitting. On the smaller Pisa dataset, the model showed superior generalization (98.93% validation accuracy vs. 98.75% for Swin-T), confirming effective handling of limited training data.

2) Contributions summary: Methodologically, this work introduced the first adaptive architecture-level fusion with learnable gating for monument recognition, unlike static ensemble methods in prior work. Empirically, we established reproducible benchmarks across two heritage datasets of varying scales. Practically, the high accuracy supports automated monument cataloging for preservation agencies, mobile tourism applications, and urban planning systems.

3) Advancement over existing methods: Unlike single-architecture approaches [9, 17, 20] that sacrifice either local or global features, and unlike static ensemble methods [15, 24] with fixed weighting, our adaptive fusion dynamically adjusts branch contributions based on input characteristics—achieving complementary feature integration previously unavailable for heritage recognition.

4) Limitations: Despite the promising results, the proposed model exhibits several limitations that warrant

acknowledgment: 1) Computational overhead with ~60M parameters against 25M for ResNet-50); 2) Evaluation limited to Egyptian/Italian heritage; 3) Fixed backbone selection not systematically optimized; 4) Limited interpretability analysis;

5) Future work: Based on our experimental findings and identified limitations, we propose the following concrete research directions: 1) Neural architecture search for optimal backbone combinations; 2) Multi-branch extensions with Vision Transformers; 3) Cross-cultural evaluation on larger heritage datasets; 4) Attention visualization for interpretability 5) Self-supervised pre-training for small datasets.

In conclusion, the proposed dual-branch fusion architecture demonstrates that architecture-level integration with adaptive feature fusion can enhance monument recognition performance, particularly on larger datasets. The approach's modular design facilitates future extensions and adaptations, providing a foundation for continued advancement in heritage monument recognition systems. We anticipate that addressing the identified limitations through the proposed future work directions will further improve the practical applicability of deep learning-based monument recognition for cultural heritage preservation.

REFERENCES

[1] E. Grilli, E. Özdemir, and F. Remondino, “Application of machine and deep learning strategies for the classification of heritage point clouds,” The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 447-454, 2019.

[2] D. Harisanty, K. L. B. Obille, N. E. V. Anna, E. Purwanti, and F. Retrialisca, "Cultural heritage preservation in the digital age, harnessing artificial intelligence for the future: a bibliometric analysis," Digital Library Perspectives, Vol. 40, No. 4, pp. 609–630, 2024.

[3] H. Parzinger, “Togetherness: A new heritage deal for Europe,” European Investment Bank, 2020.

[4] G. E. Halkos and P.-S. C. Aslanidis, "Evaluating the tangible and intangible parameters of cultural heritage: an economic meta-analysis in a global context," Discover Sustainability, Vol. 5, Article 398, 2024.

[5] S. Kavitha, S. Mohanavalli, B. Bharathi, C. H. Rahul, S. Shailesh, and K. Preethi, “Classification of Indian monument architecture styles using bi-level hybrid learning techniques,” In Inventive Systems and Control: Proceedings of ICISC 2022 (pp. 471-488). Singapore: Springer Nature Singapore, 2022.

[6] S. Jindam, J. K. Mannem, M. Nenavath, and V. Munigala, "Heritage Identification of Monuments using Deep Learning Techniques," Indian Journal of Image Processing and Recognition, Vol. 3, No. 4, pp. 1–7, June 2023.

[7] M. Ćosović, and R. Janković, “CNN classification of the cultural heritage images,” In 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH) (pp. 1-6). IEEE, March 2020.

[8] U. Kulkarni, S. M. Meena, S. V. Gurlahosur, and U. Mudengudi, “Classification of cultural heritage sites using transfer learning,” In 2019 IEEE fifth international conference on multimedia big data (BigMM) (pp. 391-397). IEEE, September 2019.

[9] A. Sasithradevi, B. Chanthini, T. Subbulakshmi, and P. Prakash, “MonuNet: a high performance deep learning network for Kolkata heritage image classification,” Heritage Science, 12(1), 1-14, 2024.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708), 2017.

[12] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, ... and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” In Proceedings of the IEEE/CVF international conference on computer vision (pp. 10012-10022), 2021.

[13] T. Boyadzhiev, G. Lagani, L. Ciampi, G. Amato, and K. Ivanova, “Comparison of Different Deep Neural Network Models in the Cultural Heritage Domain,” arXiv preprint arXiv:2504.21387, 2025.

[14] G. Amato, F. Falchi, and C. Gennaro, “Fast image classification for monument recognition,” Journal on Computing and Cultural Heritage (JOCCH), 8(4), 1-25, 2015.

[15] T. Djelliout, and H. Aliane, “Multi-CNN Model for Multi-Classification of Cultural Heritage Monuments,” In 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-5). IEEE, April 2024.

[16] T. Djelliout, and H. Aliane, “Building and evaluation of an Algerian Cultural Heritage dataset using convolutional neural networks,” In 2022 4th International Conference on Pattern Analysis and Intelligent Systems (PAIS) (pp. 1-7). IEEE, October 2022.

[17] S. Khandelwal, A. Prasad, A. Kumar, J. Gautam, and A. Patle, “A study on efficient image classification of historical monuments using CNN,” In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 220-225). IEEE, August 2023.

[18] V. Kukreja, R. Sharma, and S. Vats, “A Hybrid Deep Learning Approach for Multi-Classification of Heritage Monuments Using a Real-Phase Image Dataset,” In 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA) (pp. 29-32). IEEE, August 2023.

[19] V. Kukreja, R. Sharma, and D. Bordoloi, “Application of deep learning strategy for multi-classification of Indian heritage images,” In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE, July 2023.

[20] M. A. Hassan, A. Hamdy, and M. Nasr, “Egypt monuments dataset version 1: A scalable benchmark for image classification and monument recognition,” International Journal of Advanced Computer Science and Applications, 14(4), 2023.

[21] M. N. Razali, E. O. N. Tony, A. A. A. Ibrahim, R. Hanapi, and Z. Iswandono, “Landmark recognition model for smart tourism using lightweight deep learning and linear discriminant analysis,” International Journal of Advanced Computer Science and Applications, 14(2), 2023.

[22] E. O. Nixon and M. N. Razali, “Ums Landmark Recognition Dataset,” https://doi.org/10.34740/KAGGLE/DS/1877538, 2022.

[23] S. Lazebnik, C. Schmid, and J. Ponce, ”Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” In 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06) (Vol. 2, pp. 2169-2178). IEEE, June 2006.

[24] D. Kumar, Y. Kumar, V. Kukreja, S. Hariharan, B. Goyal, and A. Verma, “Preserving heritage palaces: A deep learning cnn-svm hybrid approach for multi-classification,” In 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE, July 2023.

[25] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, “Efficient transformers: A survey,” ACM Computing Surveys, 55(6), Article 109, 28 pages. https://doi.org/10.1145/3530811, 2022.

[26] R. Csordás, K. Irie, and J. Schmidhuber, “The devil is in the detail: Simple tricks improve systematic generalization of transformers,” arXiv preprint arXiv:2108.12284, 2021.

[27] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, “Autoaugment: Learning augmentation strategies from data,” In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 113-123), 2019.

[28] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A convnet for the 2020s,” In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976-11986), 2022.

[29] M. Tan, and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” In *International conference on machine learning* (pp. 6105-6114). PMLR, 2019.