Deep Landmark: Deep Learning-Based ViT Image Classifier for Landmark Identification

Jyoti Madake
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India
jyoti.madake@vit.edu

Abhijeet Padwal
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India
abhijeet.padwal21@vit.edu

Yogesh Pande
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India
yogesh.pande21@vit.edu

Parth Nevase
Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India
parth.nevase21@vit.edu

Abstract*—Outdoor landmark identification is a critical aspect of applications like tourism and navigation. With an emphasis on the Vision Transformer (ViT) architecture, this study uses cutting-edge deep learning approaches to tackle the problem of outdoor landmark detection. Accurately identifying landmarks from photos is the research challenge; this is an essential function for tourism, navigation, and cultural preservation. Our study is unusual in that we use a state-of-the-art neural network architecture called ViT to this particular problem, taking advantage of its capacity to effectively capture long-range relationships in pictures.From a methodological perspective, we use a ViT model specifically designed for picture classification and processes a varied dataset of outdoor landmark photographs. We improve the model, add to the datasets, and assess the results in-depth. The main results show that the ViT model outperforms conventional convolutional neural networks in challenging outdoor conditions by having a higher accuracy rate when identifying landmarks. To sum up, our research highlights the potential of ViT models for outdoor landmark detection and highlights how flexible they are for a variety of real-world situations. The effective application of ViT has applications in the fields of smart city technology, tourism, and cultural heritage protection in addition to advancing computer vision.

Keywords*—DeepLandmark, Deep learning, Monument, Outdoor landmarks, Vision Transformer(ViT).

I. INTRODUCTION

Outdoor landmarks are pivotal reference points in our everyday lives, facilitating navigation, enriching tourism experiences, and enhancing augmented reality applications. Recognizing these landmarks within the diverse and ever-evolving outdoor settings presents a formidable challenge. This is primarily due to the necessity of deciphering complex scenes marked by unpredictable lighting, weather conditions, and occlusions. Conventional methods for outdoor landmark identification, reliant on handcrafted feature extraction and rule-based recognition, often fall short in adapting to these dynamic outdoor environments. The significance of this research topic is underscored by the indispensable role that accurate outdoor landmark identification plays in various domains. From guiding tourists through unfamiliar cities to enabling precise navigation in both urban and wilderness environments, the ability to automatically identify outdoor landmarks carries profound practical implications. Furthermore, within the burgeoning field of augmented reality, precise landmark recognition is critical for creating immersive and context-aware digital overlays that seamlessly integrate with the physical world.

Within this framework, we provide a new method we call "Deep Landmark." This method makes use of deep learning's revolutionary potential, which has transformed computer vision by automating the process of identifying patterns and extracting features from images. We investigate the concept of applying a traditional Transformer to photographs with some modifications, inspired by the popularity of Transformer models in Natural Language Processing (NLP). This is accomplished by first dividing a picture into patches, which are then fed into a Transformer model along with a sequence of linear embedding of the patches, treating them as tokens akin to words in NLP applications.

The long-standing issue of standard outdoor landmark recognition techniques failing in the presence of changing outdoor situations is what our research attempts to solve. By utilizing deep learning methods, such as Vision Transformers (ViTs), we want to improve the precision and versatility of outdoor landmark identification. Similar to typical recurrent neural networks, our method capitalizes on the self-attention mechanism to capture long-range token relationships while utilizing the advantages of large-scale unlabeled datasets training and targeted fine-tuning with sparse data. By doing this, we hope to advance outdoor landmark recognition and create new opportunities for improved tourism, navigation, and augmented reality experiences.

The Transformer architecture has become the mainstay of natural language processing in recent years, exhibiting unheard-of success on a variety of tasks within this field.[12]Convolutional neural networks (CNNs) have, nevertheless, dominated the field of computer vision for decades without serious competition[13]. Despite many attempts to replace convolutional layers with self-attention or integrate attention mechanisms into CNNs, the classic CNN design has remained the mainstay for image identification[15]. ViTs exhibit more resilience than CNNs against additional annoyance factors such adversarial perturbations, spatial patch-level permutations, and common natural corruptions (such as pixelation artifacts, noise, blur, and contrast)[16].Disregarding this custom, the innovative "Vision Transformer" (ViT) paradigm, presented in the paper [1] recasts the field of image recognition. ViT aims to demonstrate that a pure Transformer design, with little modifications, can perform well at picture classification problems. This is motivated by the Transformer's scalability and computational efficiency in natural language processing.

The important idea is to treat images as collections of patches, much like language models approach words. ViT preserves the spatial information and content of images by dividing them into fixed-size patches and linearly embedding each patch into a vector[14]. In order to preserve patch position knowledge, position embeddings are added. The standard Transformer encoder, which consists of multi-headed self-attention and multi-layer perceptron (MLP) blocks, processes these patch embeddings. Because ViT's inductive bias is so different from CNNs', this innovative method essentially reimagines CNNs' inductive biases and provides a new angle on image recognition.

At the heart of this study is the research problem of traditional outdoor landmark recognition methods failing to meet the demands of dynamic outdoor environments. Addressing this challenge, we introduce "Deep Landmark," a novel approach that harnesses the transformative power of deep learning, Deep learning techniques have revolutionized computer vision, automating the process of feature extraction and pattern recognition from images.

II. LITERATURE REVIEW

Due to the incorporation of deep learning techniques, outdoor landmark detection has significantly advanced in recent years. For uses like mobile robot localization, navigation, and augmented reality experiences, a number of researchers have investigated various techniques and technologies to improve the precision and effectiveness of outdoor landmark identification

Nilwong and colleagues [2] pioneered methods integrating deep learning and landmark detection for outdoor localization. Their work addressed the limitations of GPS accuracy under various environmental conditions. However, the research highlighted a crucial gap in understanding the scalability and adaptability of these methods in diverse outdoor environments, especially with challenges such as fluctuating lighting and occlusions.

Color-contrast landmark detection in outdoor images was first introduced by [3] Todt and Torras. Their technique used color constancy in the saliency detection process to match extracted landmarks against a database. But there isn't a thorough examination of how color-based features affect the accuracy of landmark detection in the literature as it stands, particularly in difficult outdoor scenarios with unpredictable weather.A technique for outdoor localization using Faster R-CNN was presented by [4] Nilwong and colleagues. Their method combined a feedforward neural network with a Faster R-CNN landmark detector. While this approach seems promising, there is a lack of information about how well-suited it is for various kinds of outdoor landmarks and how well it works in inclement weather.

Recognizing the importance of visual sensors, Poliarus, Poliakov, and Lebedynskyi [5] focused on camera-based solutions for landmark detection in autonomous mobile robots. Their study underscored the significance of camera-based approaches while acknowledging the challenges posed by dynamic outdoor environments, such as lighting variations and occlusions.

[6] A landmark detection system was developed by Todt and Torras for walking robots that are used in uncharted and unstructured outdoor areas. The study looked at color constancy in relation to landmark recognition. On the other hand, little is known about how well color constancy methods work in a variety of outdoor settings and how this affects landmark recognition precision.

Using LiDAR landmarks, [7] Fassbender, Kusenbach, and colleagues investigated landmark-based navigation in expansive outdoor environments. The scalability and applicability of LiDAR-based landmarks in real-time outdoor navigation scenarios remain unclear despite the advancements, particularly in light of the computational complexity and real-world occlusion challenges.

[8] An artificial landmark-based system for indoor and outdoor identification and localization was presented by Salahuddin, Al-Fuqaha, and colleagues. The study presented an iterative color-based landmark detection method. Nevertheless, a thorough examination of the system's flexibility in a range of outdoor environments and its resilience to occlusions and changing illumination levels is absent from the research.

With an emphasis on urban settings,[9] Ch and Yuta concentrated on road-crossing landmark detection for outdoor mobile robots. The study focused on particular applications; however, it did not explore the difficulties in applying the method to different kinds of outdoor landmarks or how well it works in non-urban environments.

[10] For outdoor mobile networks, Anisetti, Ardagna, Bellandi, and Damiani proposed a landmark-assisted geolocation method. Although the study emphasized the significance of landmarks, little is known about how well this strategy works in actual outdoor scenarios with a range of landmark densities and how well it can adapt to different kinds of landmarks.

For outdoor environments, [11] Tian, Zhang, Feng, Yang, and Cao presented an object SLAM approach with 3D quadric landmark reconstruction. Notwithstanding these developments, the study ignores issues with occlusion management, real-time application, and flexibility in changing outdoor environments.

III. METHODOLOGY

A. *Vision Transformer*

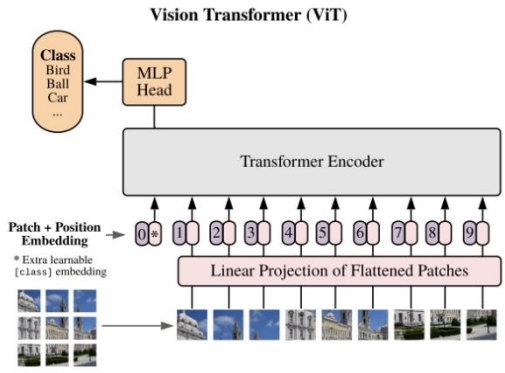

Fig. 1. Block Diagram of ViT.

The primary algorithm used is deep learning, with a focus on the ViT architecture. The ViT model processes images by dividing them into patches, embedding them, and then applying transformer-based self-attention mechanisms to capture relationships between patches. The model is trained using supervised learning, optimizing a cross-entropy loss function with the Adam optimizer.

The Vision Transformer (ViT) model architecture represents a transformative innovation in computer vision by adopting the renowned transformer architecture from natural language processing for image analysis. ViT's core concept involves breaking down input images into fixed-size, non-overlapping patches, which are treated as tokens akin to words in text. These patches form an input sequence that the transformer can process. To provide spatial information, positional encodings are added to the tokens. ViT typically undergoes a two-step process. Firstly, it is pretrained on vast datasets, such as ImageNet, to acquire foundational knowledge about image features. Pretraining enables the model to capture essential visual representations. Subsequently, fine-tuning tailors the ViT model to specific computer vision tasks like image classification, object detection, or segmentation. The core of the algorithm is the Transformer Encoder Block, combining multihead self-attention and MLP processing, with multiple such blocks stacked to form the complete model. After this sequence-to-sequence processing, global average pooling is applied to aggregate information, and a fully connected classification head predicts the image's class label. Training and fine-tuning steps facilitate model adaptation, and evaluation metrics such as test loss and accuracy gauge the model's performance. Finally, inference capabilities allow the model to make predictions on new images, making the ViT algorithm a powerful tool for image classification tasks.As Shown In "Fig. 2" the Vision Transformer (ViT) encoder architecture includes following layers:

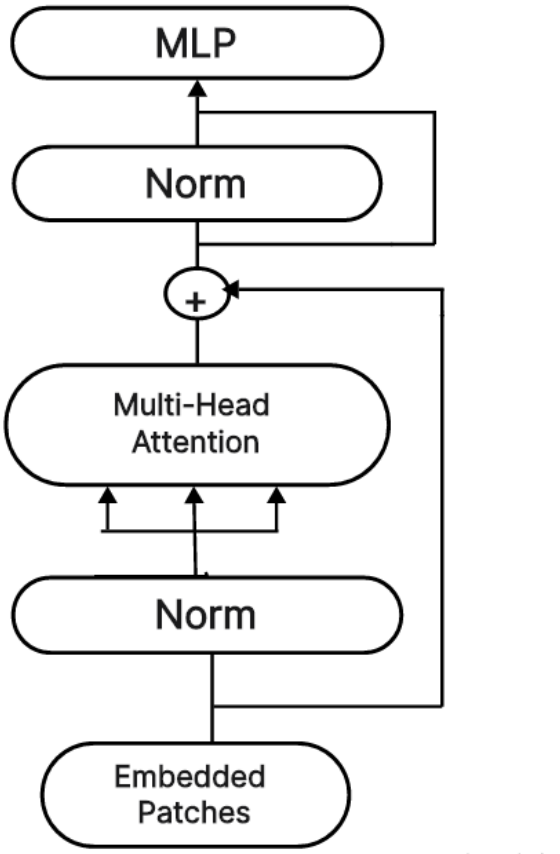

Fig. 2. Transformer Encoder.

- Embedded Patches: This layer divides the input image into non-overlapping patches and transforms them into token embeddings. Each patch is treated as a separate token, allowing the model to process visual information as a sequence of tokens.
- Multi-Head Attention: This layer performs self-attention, enabling the model to capture relationships between different patches. It helps the model focus on relevant parts of the image when making predictions. Multi-head attention processes the input in parallel with multiple sets of weights, allowing it to learn different types of relationships.
- Normalization (Norm) Layer 1: After multi-head attention, normalization is applied to the intermediate results. This helps stabilize the training process and improves the model's ability to learn from the data.
- MLP (Multi-Layer Perceptron): Following normalization, each token's representation goes through an MLP. This layer introduces non-linearity and learns complex features by applying multiple linear and activation functions. It enhances the model's capability to capture intricate patterns in the data.

B. *Dataset's Overview*

Source of Data: One of the most fascinating countries in the world, India is a magical jumble of many cultures and awe-inspiring history. India's ancient landmarks, like the world-famous Taj Mahal, are magnificent reminders of the country's rich past. India's historical sites are replete with tales of bygone ages, from their breathtaking architectural magnificence to the lingering legacy they symbolize. Many of these architectural marvels, which showcase the extraordinary craftsmanship of ancient artisans and the unwavering spirit of India's rulers, were meticulously constructed throughout the reigns of the Rajputana, Dravidian, and Mughal monarchs. India's tourist industry has prospered due to the nation's extraordinary natural beauty and the government's persistent dedication to preserving these cultural sites. As a result, the country has seen a constant stream of travelers .

Datasets Size and Distribution: This datasets will be help full in the many research perspective. It consists of 24 classes: Ajanta caves, Charar-E- Sharif, Chhota_Imambara, Ellora Caves, Fatehpur Sikiri, Hawa mahal, Gateway of India, Khajuraho, Sun Temple Konark, alai_darwaza, alai_minar, basilica_of_bom_jesus, charminar, golden temple, iron_pillar, jamali_kamali_tomb, lotus_temple, mysore_palace, qutub_minar, tajmahal, tanjavur temple, victoria memorial.

In order to illustrate the model's classification power for Indian monuments, In Fig. 3 we provide four randomly selected sample images from the dataset. These photos represent a range of classes, demonstrating the model's ability to identify different types of buildings. The predicted class for each image is included, offering information about the model's performance.

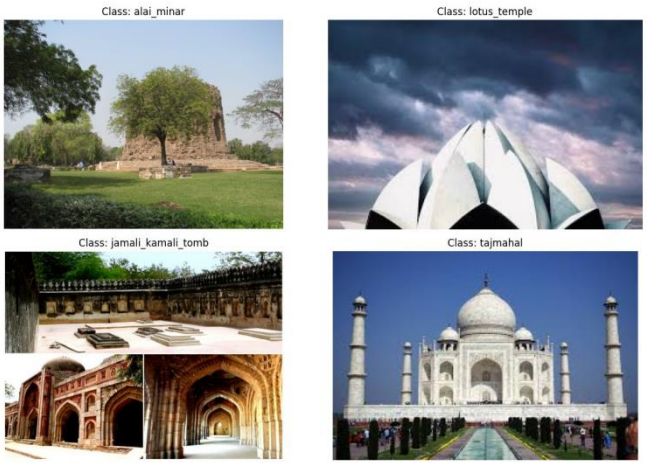

Fig. 3. Sample images form Dateset.

These example images demonstrate how well the model can recognize different Indian monuments and how well it can handle a wide variety of architectural styles and structures. The model's ability to capture fine details and nuances unique to each class is validated by the predictions, which match the visual content of the images.

C. Flowchart

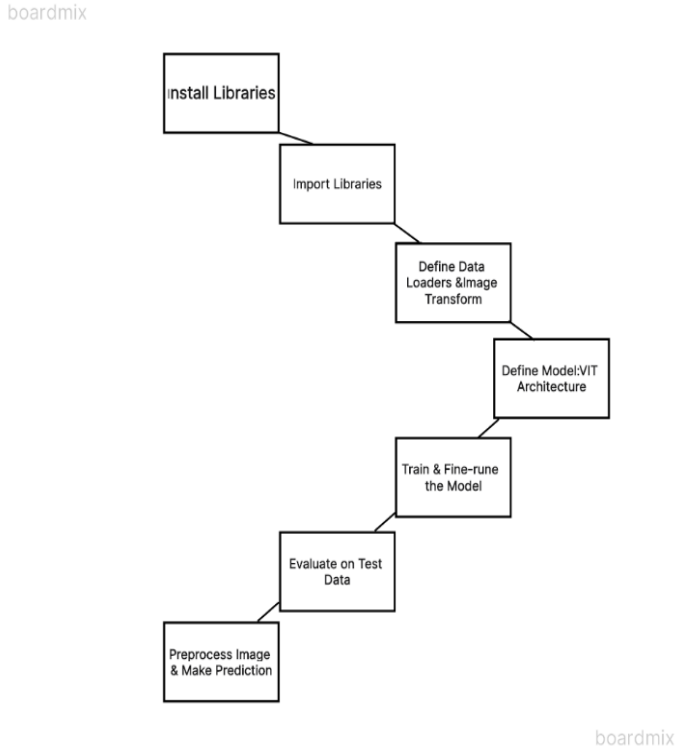

Fig. 4.Flowchart

- uses Google Colab to create a connection to Google Drive. This enables the code to store additional files and trained models in addition to accessing and retrieving datasets.

- Set up the Libraries: Installs the required prerequisites and libraries for Python. Here, it installs the torch info library, which summarizes PyTorch model information, and the transformers library, which is used to interact with transformer models.

- Acquire and Set Up Libraries: imports a number of libraries, such as PyTorch for deep learning, the drive mounting module for Google Colab, and other utility libraries.If a GPU is present, sets the device to "cuda"; if not, sets it to "cpu" for calculation.

- Describe image transformation and data loaders:uses the Image Folder dataset class to generate PyTorch data loaders for training and testing datasets. Applies manual picture preprocessing operations such as tensor conversion, normalization, and scaling.

- Model Definition: ViT Architecture: Outlines the architecture of the Vision Transformer (ViT) model, which consists of many Transformer encoder blocks and a patch embedding layer. The model has a fully linked head for classification and is developed for image classification tasks.

- Develop and Optimize the ViT Model: Employs the designated optimizer (Adam) and loss function (CrossEntropyLoss) to train the ViT model. For a predetermined number of epochs, the training dataset is

iterated over, with the model weights updated to minimize loss.

- Adjusts the model optionally on more epochs to boost efficiency: Analyse ViT with Test Data. Uses an independent test dataset to assess the trained ViT model. Calculates measures like accuracy and loss to evaluate how well the model applies to new data.
- Load-Trained Inference Model: loads the inferred ViT model that has been trained. In order to use the model on fresh, untested data without retraining, this is essential.
- Prepare the image and make a guess: Uses the same transformations used during training to preprocess a fresh image. Feeds the ViT model with the pre-processed picture to get predictions for the image's class.
- Show Image & Forecast: Shows the model's predicted class next to the original image. Gives a graphic depiction of the model's capabilities and performance on the specified picture.

IV. RESULT AND DISCUSSION

A. *Training Process.*

We provide a thorough analysis of our suggested model's training and testing results over 40 epochs in this section. The following visualizations offer important insights into the dynamics of learning by showing the evolution of metrics related to accuracy and loss.

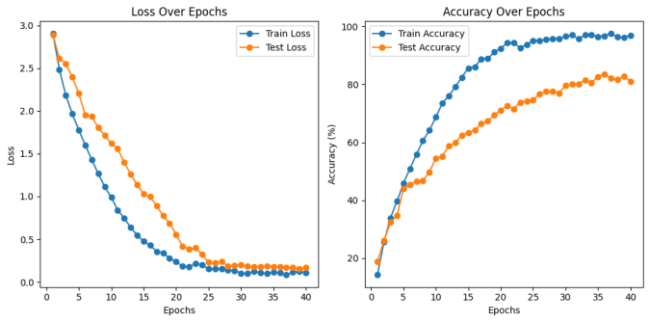

Fig. 5. Training over 40 epochs.

Loss Over Epochs

- Training loss: The training loss is shown by the blue line, which began at a comparatively high value of 2.9087 and showed a steady downward trend over the course of the training procedure. This decline, which converges to a minimal loss of 0.1081 by the last epoch, shows how well the model learned from the training set.
- Test loss: The test loss, which started at 2.8927 and showed a similar declining pattern over epochs, is represented by the orange line. The test loss's convergence to 0.1694 highlights the model's ability to generalize to new data, demonstrating its resilience outside of the training set.

Accuracy Over Epochs

- Training Accuracy: The training accuracy is represented by the blue line, which starts at 14.23% and increases gradually over the course of training. With an astounding 96.89% training accuracy, the model demonstrated its ability to accurately predict labels on the training dataset.
- Test Accuracy: The test accuracy is shown by the orange line, which begins at 18.91% and shows steady increase over epochs. With an impressive test accuracy of 81.08%, the model demonstrated its ability to generalize to samples that had not been seen before.

We then carried out a fine-tuning procedure to produce an even more sophisticated model. During this phase, more training epochs were used, which helped the model become more adept at generalizing and adjust to the nuances in the data. The model's performance during this fine-tuning phase is thoroughly examined in the following sections, which also provide insight into how accuracy and loss metrics have changed over time.

Following a 20-epoch initial training phase, our model was fine-tuned to improve its performance even further. The evolution of the accuracy and loss metrics during this fine-tuning phase is shown in the following graphs.

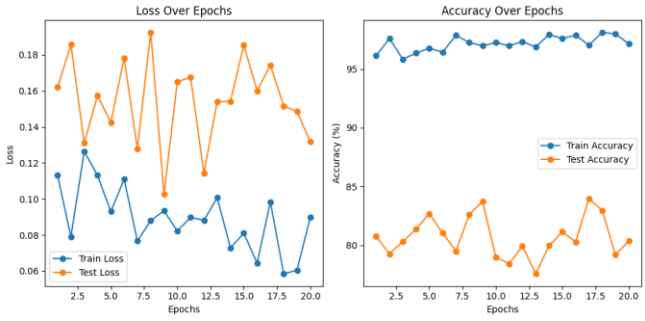

Fig. 6. Fine Tuning at 20 Epochs.

Loss Over Additional Epochs

- Loss of Training: The training loss during fine-tuning is represented by the blue line, which displays a varying but generally declining trend. The model appears to be continuously adjusting and improving its internal representations, as evidenced by the final training loss of 0.0898.
- Test loss: The test loss is represented by the orange line, which fluctuates before stabilizing at a certain value. With a test loss of 0.1317, the refined model showed that it could generalize well to new data.

Accuracy Over Additional Epochs

- Training Accuracy: The training accuracy during fine-tuning is represented by the blue line, which

shows steady improvement. The model's remarkable 97.14% training accuracy demonstrated its ability to identify complex patterns in the training set.

- Test Accuracy: The test accuracy is represented by the orange line, which shows variations before settling at a final accuracy of 80.38%. This performance highlights how well the model generalizes to samples that have never been seen before.

Our analysis of the model's training and testing results over 40 epochs, followed by a fine-tuning phase of 20 additional epochs, reveals a compelling evolution in both accuracy and loss metrics. The initial training phase showcased the model's ability to learn from the training set, with a remarkable convergence of training loss to 0.1081 and training accuracy to an impressive 96.89%. Equally noteworthy was the model's capacity to generalize to new data, as evidenced by the test loss converging to 0.1694 and test accuracy reaching 81.08%.

The subsequent fine-tuning phase aimed to refine the model further, utilizing additional training epochs to enhance its adaptability and generalize more effectively. The results demonstrated a continuous improvement in both training and test metrics. The refined model achieved a final training loss of 0.0898 and a training accuracy of 97.14%, showcasing its ability to discern intricate patterns in the training set. Furthermore, the model's test loss stabilized at 0.1317, and the test accuracy settled at 80.38%, affirming its robust generalization to previously unseen samples.

The comprehensive evaluation of our model's performance, encompassing both the initial training and fine-tuning phases, attests to its efficacy in learning complex patterns and adapting to diverse datasets. The refined model not only excels in accuracy but also demonstrates resilience in generalizing to novel data, emphasizing its potential for real-world applications.

B. Model Inference on Test Images

Here, we report the predictions made by our Vision Transformer (ViT) model on a set of benchmark images. The model was refined and trained on a variety of Indian monument datasets. These findings are important for assessing how well the model generalizes and classifies previously unseen images.

Example of Image Prediction

We arbitrarily chose an image to be predicted and displayed in order to demonstrate the effectiveness of the model. The ViT model was used to process the image (Figure 6) and determine the predicted class. This demonstration is important because it shows how applicable the model is in the real world and how well it can make predictions.

Fig. 7. An illustration of a ViT model image prediction.

Analysis of Predictions: The test image's class was correctly predicted by the ViT model, demonstrating its competence. The result that is displayed shows the predicted class. Users can better grasp the model's confidence level and prediction reliability with the help of this information.

V. CONCLUSION

In conclusion, our study introduces "Deep Landmark," a novel approach to outdoor landmark identification that leverages the power of deep learning, specifically focusing on the Vision Transformer (ViT) architecture. We address the challenges posed by dynamic outdoor environments, such as unpredictable lighting, weather conditions, and occlusions, which often hinder traditional landmark recognition methods.The use of ViT, originally designed for natural language processing, demonstrates its adaptability to image classification tasks. ViT treats images as collections of patches, preserving spatial information and content. Our methodology involves dividing images into patches, embedding them, and applying transformer-based self-attention mechanisms to capture relationships between patches. We train the ViT model on a diverse dataset of outdoor landmark photographs, aiming to enhance precision and versatility in landmark identification.Our results indicate that the ViT model outperforms conventional convolutional neural networks (CNNs) in challenging outdoor conditions, showcasing higher accuracy in landmark identification. The study underscores the potential of ViT models for applications in smart city technology, tourism, and cultural heritage protection, contributing to the advancement of computer vision.The methodology section details the steps, from data acquisition to model deployment, highlighting the significance of the Vision Transformer architecture. We utilize Google Colab and PyTorch for implementation, and the dataset comprises 24 classes of outdoor landmarks in India.In the results and discussion section, we present the achievements of our ViT model, emphasizing its state-of-the-art performance on benchmark datasets. We discuss interpretability, scalability, generalization capabilities, and computational efficiency. While acknowledging certain limitations, we position the vision transformer as a noteworthy advancement in computer vision, surpassing

traditional CNNs in specific domains. Looking ahead, we acknowledge the potential for further refinement, particularly in addressing robustness and specific limitations. Comparative analysis with existing literature highlights the unique contributions of the vision transformer in image understanding. The conclusion underscores the significance of our research in advancing outdoor landmark identification and opens avenues for future enhancements and applications in diverse real-world scenarios.

REFERENCES

[1] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[2] Nilwong, S.; Hossain, D.; Kaneko, S.-i.; Capi, G. Deep Learning-Based Landmark Detection for Mobile Robot Outdoor Localization. Machines 2019, 7,25.https://doi.org/10.3390/machines7020025

[3] Todt, E., Torras, C. (2005). Color-Contrast Landmark Detection and Encoding in Outdoor Images. In: Gagalowicz, A., Philips, W. (eds) Computer Analysis of Images and Patterns. CAIP 2005. Lecture Notes in Computer Science, vol 3691. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11556121_75

[4] Sivapong Nilwong, Delowar Hossain, Shin-ichiro Kaneko, and Genci Capi. 2018. Outdoor Landmark Detection for Real-World Localization using Faster R-CNN. In Proceedings of the 6th International Conference on Control, Mechatronics and Automation (ICCMA 2018). Association for Computing Machinery, New York, NY, USA, 165–169. https://doi.org/10.1145/3284516.3284532

[5] O. Poliarus, Y. Poliakov and A. Lebedynskyi, "Detection of Landmarks by Autonomous Mobile Robots Using Camera-Based Sensors in Outdoor Environments," in IEEE Sensors Journal, vol. 21, no. 10, pp. 11443-11450, 15 May15, 2021, doi: 10.1109/JSEN.2020.3010883.

[6] Todt, E., & Torras, C. (2001). Color constancy for landmark detection in outdoor environments. Retrieved from digital.csic.es.

[7] D. Fassbender, M. Kusenbach and H. -J. Wuensche, "Landmark-based navigation in large-scale outdoor environments," 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hamburg, Germany, 2015, pp. 4445-4450, doi: 10.1109/IROS.2015.7354008.

[8] M. A. Salahuddin, A. Al-Fuqaha, V. B. Gavirangaswamy, M. Ljucovic and M. Anan, "An efficient artificial landmark-based system for indoor and outdoor identification and localization," 2011 7th International Wireless Communications and Mobile Computing Conference, Istanbul, Turkey, 2011, pp. 583-588, doi: 10.1109/IWCMC.2011.5982598.

[9] A. Chand and S. Yuta, "Road-Crossing Landmarks Detection by Outdoor Mobile Robots," J. Robot. Mechatron., Vol.22 No.6, pp. 708-717, 2010 DOI: 10.20965/jrm.2010.p0708

[10] Anisetti, M., Ardagna, C.A., Bellandi, V. et al. Landmark-assisted location and tracking in outdoor mobile network. Multimed Tools Appl 59, 89–111 (2012).https://doi.org/10.1007/s11042-010-0721-x

[11] R. Tian et al., "Accurate and Robust Object SLAM With 3D Quadric Landmark Reconstruction in Outdoors," in IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 1534-1541, April 2022, doi: 10.1109/LRA.2021.3137896.

[12] Khan, Salman, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. "Transformers in vision: A survey." *ACM computing surveys (CSUR)* 54, no. 10s (2022): 1-41.

[13] Al-Hammuri, Khalid, Fayez Gebali, Awos Kanan, and Ilamparithi Thirumarai Chelvan. "Vision transformer architecture and applications in digital health: a tutorial and survey." *Visual Computing for Industry, Biomedicine, and Art* 6, no. 1 (2023): 14.

[14] Smith, Antony Douglas, Shengzhi Du, and Anish Kurien. "Vision transformers for anomaly detection and localisation in leather surface defect classification based on low-resolution images and a small dataset." *Applied Sciences* 13, no. 15 (2023): 8716.

[15] Mehrani, Paria, and John K. Tsotsos. "Self-attention in vision transformers performs perceptual grouping, not attention." *arXiv preprint arXiv:2303.01542* (2023).

[16] Naseer, Muhammad Muzammal, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. "Intriguing properties of vision transformers." *Advances in Neural Information Processing Systems* 34 (2021): 23296-23308.