International Journal of Informatics and Computation (IJICOM)
Vol. 7, No. 2, 2025
ISSN: 2685-8711, E-ISSN: 2714-5263 DOI: 10.35842/ijicom

Identifying Traditional Malay Building Architectural Styles Using Vision Transformer Architecture

Heri Pramono[1], Sri Winiarti[2], Abdul Fadlil[3], Sunardi[4]

Abstract

The preservation and documentation of traditional Malay buildings is a significant challenge, especially in identifying diverse architectural styles, which is often done manually. This study aims to optimize digital architecture using Vision Transformer (ViT) for identifying Malay architectural styles, such as Riau Malay and Pontianak Malay, by measuring model performance using Precision, Recall, and F1-Score. The method used is ViT-based deep learning trained using a dataset of traditional building images. The data was divided using an 80:20 and 70:30 ratio for training and testing data. The model was optimized to improve accuracy and prevent overfitting using regularization techniques. Testing results show that the ViT model achieved excellent Precision, Recall, and F1-Score values, with Precision and Recall reaching 0.99 on the training data, and 0.98 for Riau Malay House Types and 0.97 for Pontianak Malay Traditional Houses on the test data. This proves that ViT can automatically and accurately identify Malay architectural styles. This research has the potential to be applied in digital preservation, traditional building documentation, and the development of AI-based applications for the cultural and tourism sectors.

Keywords:
Automatic Identification, Traditional Malay Houses, Vision Transformer, Deep Learning

1. Introduction

Traditional Malay buildings in Indonesia and Malaysia preserve a strong cultural heritage, visual identity, and local philosophy, characterized by architectural elements such as stilted structures, carved wooden ornaments, cross ventilation systems, and open space compositions that reflect the tropical climate and local cultural values [1]. However, from a conservation perspective, significant challenges arise in relation to modernization, urbanization, and the physical degradation of traditional Malay buildings, both in Indonesia and Malaysia, which have caused many original buildings to suffer damage or lose their identity.

Corresponding Author: Sri Winiart, Universitas Ahmad Dahlani (sri.winiarti@tif.uad.ac.id)
1 Heri Pramono, Universitas Ahmad Dahlan
2 Sri Winiarti, Universitas Ahmad Dahlan
3 Abdul Fadlil, Universitas Ahmad Dahlan
4 Sunardi Universitas Ahmad Dahlan

In the context of Malaysia, traditional Malay buildings emphasize the need for global structural evaluation and greater attention to mechanical aspects, and the environmental impact on traditional Malay buildings [2].

Meanwhile, in Indonesia, the application of neo-vernacular Malay style in contemporary buildings has begun to be studied, for example, in the application of neo-vernacular Malay architecture in the North Sumatra Sports Building, which adapts traditional elements into a modern context [3]. Studies of traditional Malay buildings in Indonesia and Malaysia reveal similarities and differences in architectural style. Similarities include stilted houses, the use of local wood, pyramid or long-pitched roofs, and decorative carvings with symbolic flora and fauna motifs. Both traditions also emphasize adaptation to the tropical climate with cross ventilation and stilted structures to overcome humidity and flooding. However, there are significant differences: in Indonesia (for example, in Riau and Sumatra), Malay houses feature more detailed wooden carvings and local philosophies such as bamboo shoot motifs, while in Malaysia, colonial and Islamic influences are more pronounced, both in the shape of the roof structure, the use of interior space, and the layout of the building. [2]; [4]. In modern architectural studies, these differences are considered to be a wide variety of styles that shape the transnational identity of Malay architecture. However, these subtle visual differences pose a challenge in the process of automatic documentation and classification, as digital systems must be sensitive enough to recognize both commonalities and differences in detail within each sub-style [5];[6];[7]. Other studies that adapt traditional Malay house designs for occupant comfort also show that elements such as ventilation, room layout, and raised floor height are still relevant to apply in modern designs for better thermal comfort and quality of life [8]. In the field of contemporary architecture, the study of traditional Malay architectural design as an adaptive element in modern architecture shows that the transformation of traditional visual elements into new forms (digital architecture) is one strategy for preserving cultural identity [4].

In the field of digital architecture, visual representation and 2D imaging technologies, 3D modeling, and virtual tours have become the primary media for documenting traditional buildings. However, the identification of architectural styles still often relies on human observation, which tends to be subjective, slow, and inconsistent. Therefore, automated approaches based on artificial intelligence, particularly deep learning, are very attractive due to their ability to recognize complex visual patterns [4];[9];[10];[11].

In deep learning, conventional models such as Convolutional Neural Networks (CNNs) have been widely used for image classification and architectural pattern recognition. However, CNNs work based on local convolution operations, so they have limitations in capturing global spatial relationships between image parts. To overcome these limitations, the Vision Transformer (ViT) architecture was developed, which uses a self-attention mechanism, allowing the model to process images as a collection of patches and learn the relationships between patches globally [12];[13]. In another context, the application of Deep Learning to improve the understanding of architectural styles and eras through building facades has also been discovered [14]; [15];[16].

Research discussing technology and visual computing, the Vision Transformer (ViT) model and its variants have become popular solutions in image processing, thanks to their self-attention capabilities, which are effective in capturing global and local contexts in complex images [17].

A recent review of the Vision Transformer architecture and its variations highlights the model's ability to excel at visual recognition tasks compared to traditional CNN architectures [18]. This research introduces a global contextual mechanism to improve efficiency and spatial representation in computer vision. In the realm of practical applications, the study “A modified vision transformer framework for image-based land cover segmentation” shows that ViT modifications combined with optimization algorithms (such as the firefly algorithm) can be applied in the context of rural architectural design and

spatial planning [19]. In addition, other Vision Transformer architectural innovations, such as SpectFormer (combining multi-headed attention and spectral components), strengthen image representation capabilities more efficiently [20]. Variants such as Reversible Vision Transformers also offer reduced memory usage without sacrificing accuracy, which is relevant when architectural image data is large in size [21].

Although the potential of Vision Transformer is enormous, its specific application for identifying traditional Malay architectural styles is still limited. Several obstacles that need to be considered include:

1. Limitations of the dataset in the form of videos of traditional Malay architecture in terms of quantity, variety of angles, and regional style variations.
2. The variability of Malay sub-styles (in Indonesia: Riau and Kalimantan, while in Malaysia Negeri Sembilan and Terengganu) that produce subtle visual differences in ornamentation, roof shapes, and room details requires the model to be highly sensitive to subtle characteristics between variants.
3. The need for digital architecture optimization (e.g., image augmentation, patch embedding, fine-tuning, hyperparameter adjustment) so that the Vision Transformer model can adapt to the local cultural domain rather than general datasets such as ImageNet.
4. The issue of model interpretability and generalization must not only be accurate, but also explainable (which visual features led to style classification?) so that the results can be used in the context of cultural preservation.

Thus, this study aims to optimize Vision Transformer-based digital architecture to detect, identify, and compare the architectural styles of traditional Malay buildings in Indonesia and Malaysia. This work is expected to produce a more accurate style classification method, a rich digital database of cultural heritage, and support the process of preserving architectural heritage in the realm of digital heritage. The reason for using Vision Transformer as an optimization effort in identifying traditional Malay buildings based on digital architecture is that Vision Transformer works by breaking images into small pieces (image patches) and processing them through a self-attention mechanism. This mechanism allows the model to understand the relationship between image parts comprehensively (global context), while remaining sensitive to visual details. Recent research shows that ViT outperforms CNN in architectural classification tasks, heritage building damage detection, and facade segmentation [12]; [19];[13]. This proves that ViT has great potential in identifying complex architectural visual patterns, including Malay ornaments that are rich in detail.

2. Related Works

Research related to the recognition of traditional architectural styles through computer vision and deep learning can be grouped into several themes: (1) classification of architectural styles or facades; (2) modern architectural models (Transformer, hybrid) for architectural images; and (3) efficiency and cross-domain generalization.

In the context of architectural style, several recent studies have demonstrated the application of deep learning methods for classifying building or facade styles. For example, the study “An Approach with Deep Convolutional Neural Networks for Accurate Architectural Style Classification” (2024) proposes a deep CNN architecture optimized to distinguish architectural styles based on real-world facade features. Additionally, in the domain of urban facades [22], introducing the Revision-based Transformer Facade Parsing pipeline, which combines Vision Transformer with a line integration-based revision algorithm, demonstrating that ViT can successfully handle parsing highly complex real facades [23].

Meanwhile, the study “Architectural Style Classification Based on Deep Learning” (2025) shows that trained deep learning models can effectively extract architectural features from CAD models and images for style classification [11]. Although not specific to Malay architecture, this approach is relevant as a comparative methodology.

In the context of Vision Transformers, in recent years, vision-based transformers (Vision Transformers, ViT) and their efficient variants have been increasingly used due to their ability to model global relationships between image patches (self-attention). A survey study titled “A Comprehensive Study Showcasing Vision Transformers” (2023) compares various

Vision Transformer models across different vision tasks, highlighting the advantages of ViT in handling global features compared to traditional CNNs [24]. Another review, "Understanding the architecture of vision transformer and its variants: A review" (2024), presents an in-depth analysis of ViT variants, including their advantages and limitations in the context of images [18].

However, in the domain of multiscale vision and efficiency, several efficient architectures have emerged. For example, TurboViT uses generative architecture search to produce a lighter ViT design that remains accurate [25]. This approach is relevant when field implementation (mobile/web) requirements are a consideration. Also, the big.LITTLE Vision Transformer architecture proposed by Guo et al. uses dual blocks (large blocks + efficiency blocks) with dynamic inference to balance accuracy and computational efficiency [25]. Furthermore, hybrid models that combine CNN and Transformer have gained attention. Chibuike et al. in "Convolutional Neural Network–Vision Transformer" present a hybrid architecture that utilizes the strong local features of CNN and the global context of ViT [26]. This approach is suitable for traditional architectural domains where local details (carvings, ornamental elements) and global structures (roof silhouettes, floor plans) are equally important.

3. Proposed Method

This research methodology is designed to optimize the application of Vision Transformer (ViT)-based deep learning in identifying patterns and styles of traditional Malay architecture in Indonesia and Malaysia. The stages of the methodology consist of:

a. **Literature Study and Needs Analysis**

This literature study and needs analysis aims to examine the theory of traditional Malay architecture in Indonesia and Malaysia, including differences and similarities in styles such as stage structures, roof shapes, ornaments, and spatial arrangements. A review of previous research on deep learning and Vision Transformer in architectural classification was also conducted to evaluate the application of technology in identifying cultural heritage. Based on this study, the requirements for an automatic identification system based on digital architectural images are formulated. The objectives of this literature study and analysis are to gain an in-depth understanding of the characteristics of traditional Malay architecture, identify relevant technologies in the field of digital architecture, and formulate effective system requirements for digitizing and classifying cultural heritage buildings efficiently and accurately.

b. **Dataset Development**

Data collection (dataset development) was carried out by gathering digital images of traditional Malay buildings from various sources, such as field documentation, academic repositories, heritage publications, and open online sources. This dataset includes variations in shooting angles, lighting conditions, and regional style differences (Sumatra, Riau, Kalimantan, Negeri Sembilan, Terengganu, etc.). The data was then labeled with architectural style categories based on expert literature. The purpose of this data collection was to build a representative and diverse dataset, which will be used in the development of an image-based automatic identification system, as well as to ensure the accuracy of traditional Malay architectural classification based on different styles and conditions.

c. **Data Preprocessing**
Data preprocessing is performed by normalizing image sizes to make them uniform (e.g., 224×224 pixels), augmenting data to increase diversity (such as rotation, flipping, and contrast/light changes), and cleaning the dataset to remove duplicate images or noise. The purpose of this activity is to prepare high-quality and varied data so that the model can be trained more optimally and produce accurate classifications.

d. **Model Development**
Deep learning modeling was performed by constructing a Vision Transformer (ViT) architecture using patch embedding and self-attention schemes, and comparing its performance with baseline models such as CNN (ResNet, EfficientNet). Fine-tuning was performed on the ViT model using the Malay architecture dataset, and the Explainable AI (XAI) method was applied to highlight the image parts that formed the basis of the classification decision. The purpose of this activity was to develop an effective model for classifying Malay architecture, as well as to ensure the transparency and interpretability of the model through XAI.

e. **Model Evaluation**
Experiments and evaluations were conducted using a train-validation-test split scheme (e.g., 70:15:15), and model performance was measured using metrics such as accuracy, precision, recall, F1-score, and confusion matrix to evaluate classification errors between sub-styles. The ViT results were compared with the CNN model as a baseline. The purpose of this activity is to evaluate the effectiveness of the model in classifying Malay architecture and to identify areas that need improvement to enhance performance.

f. **Model Optimization**
Model optimization was performed by testing the effects of hyperparameters such as learning rate, batch size, number of patches, and number of Transformer layers. In addition, transfer learning was used with pretrained ViT on a general dataset (ImageNet), which was then fine-tuned on the Malay architecture dataset. Regularization, such as dropout and weight decay, was applied to avoid overfitting. The purpose of this activity was to improve model performance, ensure good generalization, and reduce the risk of overfitting on the Malay architecture dataset.

g. **Implementation in Digital Architecture**
Implementation in digital architecture is carried out by integrating the model into a web-based system or a simple application for identifying Malay architectural styles. This system also displays classification results along with visualizations of important features, such as heat maps, to improve interpretability. The purpose of this activity is to provide a practical and easy-to-use platform for identifying and understanding Malay architectural styles, as well as ensuring transparency in the classification decisions made by the model.

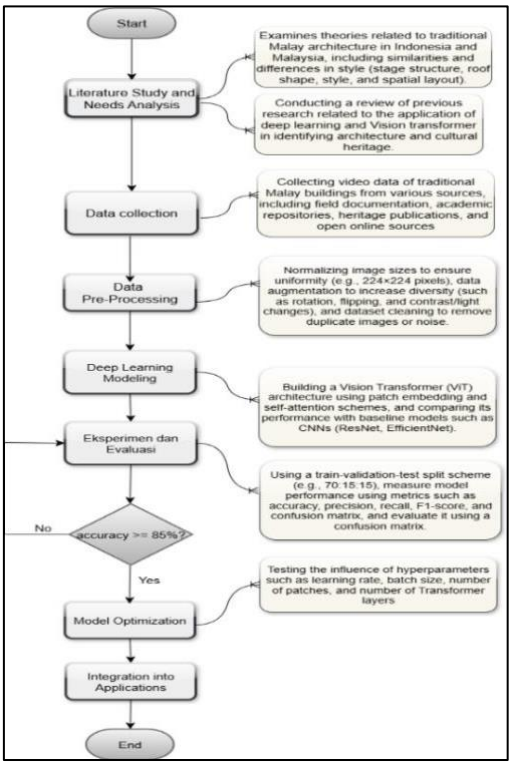

Fig 1. Research Methods

In the process of identifying these traditional Malay buildings, an artificial intelligence approach was used with Vision Transformer architecture.

A hybrid approach was developed to combine the strengths of local feature extraction from CNN with the global context of ViT. CNN extracts texture patterns and small details, while ViT models the relationships between parts of an image. Mathematically, the process can be described as:

1. Local feature extraction (CNN):

$$F_{CNN} = f_{CNN}(X) \quad (1)$$

where X is the input image $f_{CNN}(X)$ is a convolution function:

$$f_{CNN}(X) = \sigma (W * X + b) \quad (2)$$

where $*$ is the convolution operation, and σ is the ReLU activation function.

2. Projection to the ViT patch:

$$Z = \text{PatchEmbed } F_{CNN} \quad (3)$$

3. Implementation self-attention (ViT Encoder):

$$Z' = \text{TransformerEncoder}(Z) \quad (4)$$

4. Final classification:

$$y = \text{Softmax } (W_0 Z'_{cls} + b_0) \quad (5)$$

Research by Chibuike et al [26] shows that the hybrid CNN–ViT architecture produces a 3–5% increase in accuracy compared to single models, especially on complex textured datasets such as architectural carvings.

The Vision Transformer (ViT) model changes the image classification paradigm by replacing convolution operations with a self-attention mechanism capable of capturing global relationships between patches. Input images of size H x W x C are divided into N patches of size P x P so that [25]

$$N = \frac{HW}{P2} \quad (6)$$

Each patch is flattened and projected onto a D-dimensional vector through linear embedding:

$$Z_0 = [x_1E; x_2E; \dots x_NE] + E_{pos} \quad (7)$$

where E is the projection matrix and E_{pos} is the positional encoding. The main stage of ViT is the Multi-Head Self-Attention (MHSA) block, which is mathematically formulated as follows [27]:

With

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

$$\text{Q} = XW_Q = K = XW_K, V = XW_V \quad (9)$$

as query, key, and value matrices.

Each head learns the inter-patch relationship from a different perspective, then the results are combined (concatenated) and reprojected:

$$\text{MultiHead(Q, K, V)} = \text{Concat(head}_1, \dots, \text{head}_h)W_0 \quad (10)$$

Output from the Transformer Encoder block through the Layer Normalization (LN) and two-layer Feed-Forward Network (FFN) stages:

$$\text{FFN(x)} = \text{GELU}\ (xW_1 + b_1)\ W_2 + b_2 \quad (11)$$

and wrapped with a residual connection:

$$x' = x + \text{MHSA(LN(x))}, x'' = x' + \text{FFN(LN(x}')) \quad (12)$$

Finally, the [CLS] vector is used for classification through the MLP head layer with the softmax activation function [28]. This approach enables ViT to understand the global context, such as the relationship between roof shapes, wall proportions, and typical Malay building ornaments. According to Wong et al [25]. The TurboViT model utilizes generative architecture search to construct a lighter ViT without losing accuracy, enabling implementation on edge devices such as web and mobile applications.

4. Experimental Setup

A. Case Study: Traditional Malay Houses in Riau and Pontianak

Pekanbaru, as the capital of Riau Province, is not only known as an economic and trade center, but also as a silent witness to the rich development of Malay culture. One aspect that characterizes Malay culture in Pekanbaru is the traditional Malay house, which is still preserved and maintained. This building not only has aesthetic value but also contains deep philosophy, history, and local wisdom. One of the most well-known examples of traditional buildings in Pekanbaru is the Riau Malay Traditional House, which is often seen in cultural events and local community activities. The Riau Malay Traditional House is a rich cultural heritage with traditional values, combining beauty and functionality that reflects the Malay people's philosophy of living in harmony with nature. This house is an adaptation of a stilt house built high to protect its inhabitants from floods and wild animals, closely related to the geographical conditions and climate around the rivers and lowlands of Riau. The design of this house contains many symbols, such as a pointed roof that symbolizes prayers for a blessed family life, as well as open and interconnected spaces to emphasize communication and openness among family members.

Traditional Riau Malay houses are generally built with wood as the main material, often using strong and durable types of wood such as meranti or ulin, based on the tradition of stilt houses (raised) to cope with flooding and improve ventilation [29];[30]. The distinctive shape of the roof is a pyramid (or truncated pyramid), as well as other roof shapes such as the lontik roof, which curves at the ends to resemble buffalo horns or lancang. These roofs

They are not only aesthetically pleasing but also serve to reduce heat and facilitate air circulation [31]. Wooden ornaments and carvings are prominent elements; plant, animal, and even geometric motifs adorn the pillars, doors, windows, and upper edges of the roof (such as the eaves or gables). In addition, traditional houses always have wooden stairs as the entrance, which are usually located at the front; in some types of traditional houses, the number of steps is odd, with a specific philosophy behind it. The front porch is also an important part of the house structure, serving as a place to welcome guests, socialize, and as a transitional space between the outside and inside of the traditional house [31]. Figure 2 shows several types of traditional Malay houses in Riau.

Fig 2. The Shape of Traditional Malay Houses in Riau

Another traditional Malay house that has been widely discussed since its establishment is the Pontianak traditional Malay house. Pontianak, the capital of West Kalimantan Province, has a rich Malay culture that reflects the identity of its people. One of the important parts of the Malay cultural heritage in Pontianak is the Pontianak traditional Malay building. This traditional building not only has high aesthetic value but also contains deep philosophical and cultural meanings, reflecting the social, political, and historical life of the Malay community in Pontianak. Through this case study, we will explore the characteristics of the Pontianak Malay traditional building and how it symbolizes the identity and local wisdom of the Malay community in West Kalimantan. The modern traditional Malay building in Pontianak, known as the Rumah Adat Melayu Pontianak (Traditional Malay House of Pontianak) in the Jalan Sutan Syahrir Cultural Village Complex, began construction on May 17, 2003, and was completed and inaugurated on November 9, 2005, by the Vice President of the Republic of Indonesia, Jusuf Kalla, as a symbol of the cultural wealth of the Malay community in West Kalimantan [32].

The stilt house with tall pillars was built not only as an adaptation to tropical climatic conditions and flooding, but also as a public space for community activities such as deliberations, cultural gatherings, and social events, reflecting the philosophy of brotherhood, cooperation, social solidarity, and the concept of a community that lives in “shared destiny” [33]. The ornaments that adorn the building, the functional roof shape, and the layout of the public space around the building emphasize the harmonious relationship between humans, customs, and the natural environment, as well as showing the social position and cultural identity of the Pontianak Malays, which is rooted in local traditions, religion, and cultural acculturation. The traditional houses of the Pontianak Malay have distinctive features that differentiate them from other traditional Malay houses, including pointed pyramid roofs made of natural materials such as palm fiber or thatch, which serve to maintain coolness and protect from heavy rain. These buildings generally use the concept of stilt houses, built on poles to avoid flooding and provide coolness, as well as reflecting the value of togetherness in the community. In addition, traditional Pontianak houses are decorated with wood carvings depicting traditional Malay art, with geometric, floral, and fauna motifs that have spiritual meanings. The large stairs connecting the house

The ground is also often decorated with colorful carvings and symbolizes the transition from the outside world to the more sacred world inside the house.

Fig 3. Several types of traditional Malay houses in Pontianak

Based on this explanation, there are several differences and similarities between traditional Malay buildings in Riau and Pontianak in terms of location, roof shape, building structure, carvings and ornaments, social function, design philosophy, and role in society. This is shown in Table 1.

Table 1. Differences between the traditional houses of the Riau Malay and the Pontianak Malay

Aspect	Traditional Malay House of Riau	Traditional Malay House of Pontianak
Location	Province Riau, Indonesia	West Kalimantan Province
Roof Shape	Cut pyramid roof, lontik roof, kajang folding roof, peaked lontik roof	Pyramid roof, pointed roof, curved roof
Building Structure	Stilt house with tall pillars, made of wood and bamboo	Stilt house with tall pillars, made of strong wood
Carvings and Ornaments	Carvings of leaves, wings, and other decorations with symbolic meanings	Typical Malay ornaments with Javanese cultural influences
Social Function	Place of residence, traditional deliberations, and traditional ceremonies	Malay residence and cultural center
Design Philosophy	Harmony with nature, openness among family members	cooperation, sharing the same fate, and social solidarity
Role in Society	Symbols of social status and cultural identity of the Riau Malays	Symbols of cultural identity of the Pontianak Malays

In identifying the second object of these two traditional Malay buildings, a literature review was conducted by searching for image sources using conventional/manual methods. This data may be incorrect in terms of the names given by the authors; for example, traditional Malay houses in Riau may be referred to as traditional Malay houses in Pontianak, and vice versa. The development of smartphone technology with various social media applications has led to a large amount of video data being uploaded by the public, including video data of traditional Malay buildings. However, the accuracy of the information uploaded in identifying the characteristics of traditional Riau Malay houses and traditional Pontianak houses may be incorrect or invalid. There is another way that can be used to identify the architectural styles of traditional Malay houses in Riau and Pontianak by utilizing artificial intelligence. One of the methods used to recognize an object in the concept of artificial intelligence (AI) is deep learning. One of the techniques used to identify objects in the form of temporal data (or video-type data) is Vision Transformer (ViT) [34];[18].

B. Application of Vision Transformer in Identifying the Architectural Style of Traditional Malay Houses

Vision Transformer (ViT) is a deep learning model that adapts the transformer architecture from Natural Language Processing (NLP) for Computer Vision (CV) tasks. Unlike Convolutional Neural Networks (CNNs) that use local convolutions to process images, ViT divides images into small patches, flattens them, and treats them as a sequence of tokens, similar to words in NLP. Each patch is processed through a self-attention mechanism that allows the model to capture global relationships between parts of the image [35]. ViT demonstrates strong capabilities in capturing global context and long-range dependencies in images, which are often difficult for CNNs to achieve. However, ViT requires large datasets for training to compete with CNNs in terms of accuracy. To address this shortcoming, approaches such as Data-efficient Image Transformer (DeiT) have been developed, which enable training ViT with smaller datasets without sacrificing performance [18]. ViT has been applied in various domains, including object recognition, image segmentation, and medical image processing, demonstrating great potential in improving the accuracy of medical diagnoses [36]. Overall, Vision Transformer represents a significant leap forward in architectural approaches to computer vision tasks, with a focus on global processing and flexibility in handling various types of visual data [37].

This study provides an application of how automation can be used to identify the architectural styles of traditional Malay houses in Riau and Pontianak using the vision transformer model. Vision Transformer (ViT) can be used to identify architectural styles in traditional Malay houses in Riau and Kalimantan in a very effective manner, thanks to its ability to understand complex visual patterns and capture the relationships between elements in images. The basic principle of ViT is an adaptation of the transformer architecture originally used in natural language processing (NLP), converting images into a series of “patches” or small pieces. Each patch is treated like a token in NLP, which is then analyzed collectively by the model to capture the global relationships and context between the elements of the image. This differs from Convolutional Neural Networks (CNNs), which focus more on the local analysis of image features using convolution operations. ViT's process of identifying the architectural styles of traditional Malay houses in Riau and Pontianak begins with processing video data that is converted into frame-by-frame images. This conversion process in ViT is broken down into patch by patches. To begin the architectural style identification process, images of traditional Malay houses in Riau and Kalimantan will be broken down into small patches. For example, images of house facades, roof details, house pillars, etc. ViT then treats each patch as a token that can be learned. After the image is divided into patches, each patch is processed through a self-attention mechanism.

In the context of traditional houses, this allows the model to recognize design patterns in each part of the image (such as pyramid-shaped roofs, wood carvings, and stilted house structures). Self-attention helps ViT capture global information, for example, how the pointed roof design of traditional Riau Malay houses relates to the philosophy of balance and blessings that they seek to convey. The ViT model was trained using a large dataset that includes images of various types of Malay traditional houses, including Riau Malay Traditional Houses and Kalimantan Malay Traditional Houses. In this training, the model will learn the architectural elements that characterize both traditional houses, such as tall pillars on stilt houses, pyramid or pointed roofs, wood carvings, and stilt structures. ViT will also study the different design variations between traditional houses in Riau and Kalimantan, such as distinctive carving motifs and the use of local materials such as wood and bamboo. After training, ViT can be used to identify architectural styles based on images of traditional houses provided. For example, when given an image of a traditional Malay house in Riau, ViT can recognize distinctive features such as a pyramid roof made of palm fiber or thatch, a stilted structure that raises the house above the ground, and carvings rich

in symbolism. Similarly, for traditional Malay houses in Kalimantan, ViT can recognize other distinctive features such as typical Dayak carvings or the use of more diverse materials. This explanation is summarized in the form of a flowchart of the identification process for traditional Malay houses in Riau and Kalimantan, as shown in Figure 4.

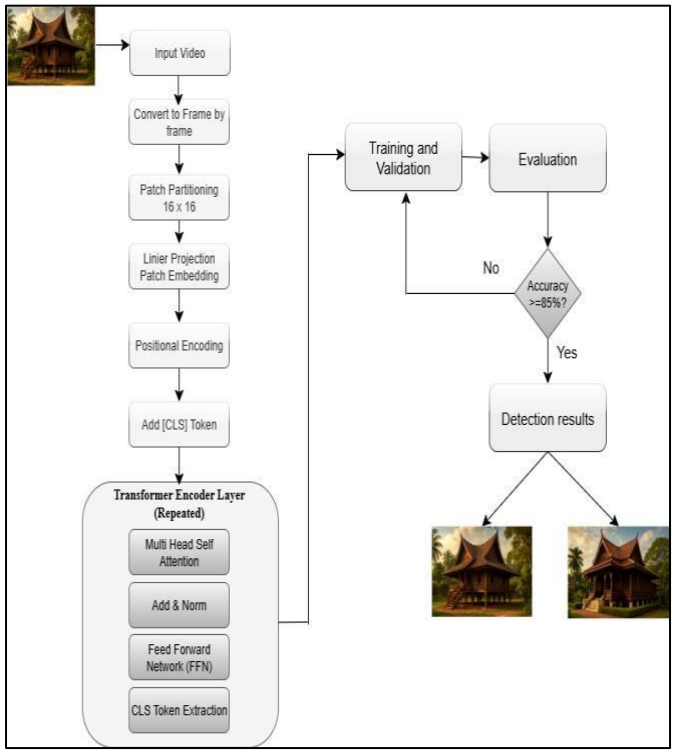

Fig 4. Application of Vision Transformer in Identifying Traditional Malay Houses

C. Data Analysis

From the survey conducted directly in Pontianak and Pekanbaru, the data were mapped into seven data groups: overall image of the building, roof, pillars, walls, doors, windows, and columns. Table 2 shows the number of frame images obtained from 5 videos for each aspect. The aspects used to identify traditional houses are five characteristics: the entire building, the shape of the roof, the structure of the building, carvings and ornaments, the shape of the windows, and the shape of the stairs, with a maximum duration of 2 minutes. Every 2 minutes produces 170 image frames.

Table 3. Data on Traditional Malay Houses in Riau

Aspect	Video duration	Number of videos	Number of frame images
Roof Shape	2 minutes	5	159 frame x 5 = 759 frame
Building Structure	2 minutes and 20 seconds	5	167 frame x 5 = 835 frame
Carvings/Ornaments	1 minute and 25 seconds	5	135 frame x 5 = 675 frame
Entire Building	2 minutes and 30 seconds	5	177 frame x 5 = 885 frame
Window Shape	1 minute and 30 seconds	5	132 frame x 5 = 660 frame
Staircase Shape	1 minute and 15 seconds	5	120 frame x 5 = 600 frame
		Total	4339 frames

Table 4. Data on Traditional Malay Houses in Riau

Aspect	Video duration	Number of videos	Number of frame images
Roof Shape	2 minutes and 10 seconds	5	161 frame x 5 = 805 frame
Building Structure	2 minutes and 30 seconds	5	177 frame x 5 = 885 frame
Carvings/Ornaments	1 minute and 15 seconds	5	120 frame x 5 = 600 frame
Entire Building	2 minutes and 45 seconds	5	198 frame x 5 = 990 frame
Window Shape	1 minute and 20 seconds	5	128 frame x 5 = 640 frame
Staircase Shape	1 minute and 20 seconds	5	120 frame x 5 = 600 frame
		Total	4520 frames

Based on Table 2, the data will be trained and validated by dividing it into an 80:20 split. In the Vision Transformer (ViT) model training process, an 80:20 split is one of the most commonly used divisions in machine learning, where 80% of the data is used for model training, and the remaining 20% is used for testing. In the context of traditional Malay houses in Riau and Kalimantan, the following is an analysis of how the 80:20 split can affect the ViT model training and evaluation process. From Table 2 and Table 3, the total image frames for each aspect of traditional houses are as follows:

a. Traditional Malay Houses in Riau: 4339 image frames
b. Traditional Malay Houses in Kalimantan: 4520 image frames

If using an 80:20 split, the data used for training and testing is described in Table 4.

Table 4. Split the Data for training and testing

Malay Traditional House Class	Total data	Training Data Amount 80%	Testing Data Amount 20%
Traditional Malay House of Riau	4339	3471	868
Traditional Malay House of Pontianak	4520	3616	904
Total	8859	7087	1772

D. Training Process

During the training phase, the Vision Transformer model will learn patterns in the training data (80% of the total image frames). ViT will break each image frame into several small patches and process the information through a self-attention mechanism to recognize existing patterns, such as roof shapes, building structures, carvings and ornaments, window shapes, and stairs. With 80% of the training data, the model will perform iterations (epochs) to minimize the loss function and improve the model's predictive capabilities.

a) Image Frame Processing: Each image frame obtained from the video will be converted into patches, and the model will use a transformer encoder to understand the relationships between patches.
b) Training: During training, the model will attempt to recognize the characteristics of traditional house architecture through elements such as roof shape, building structure, and distinctive ornaments.

The ViT model will also learn the global context of the image, such as the relationship between larger elements (e.g., the relationship between the roof and the building structure) that conventional CNN models cannot easily capture.

E. Testing Process

After the model has been trained using 80% of the training data, it will be tested using the remaining 20% of data that has been set aside for testing. The model has never seen this data during the training process, so it can provide an overview of how the model works on data that has not been seen before.

1) Accuracy Evaluation: Testing is conducted to evaluate the accuracy of model predictions. The model will be tested to identify elements of traditional house architecture (such as roof shape, stilt house structure, carvings, etc.) in previously unseen test images.
2) Evaluation Metrics: Several metrics can be used to evaluate the performance of the model, namely:
   a. Accuracy: The proportion of correct predictions compared to the total number of predictions.
   b. Precision: Measures the proportion of correct positive predictions.
   c. Recall: Measures how much positive data is successfully recognized by the model.
   d. F1-Score: The harmonic mean between precision and recall.

The purpose of testing in the context of machine learning, including when using models such as Vision Transformer (ViT) to identify architectural styles in traditional Malay houses in Riau and Kalimantan, has several very important aspects. Testing is used to measure the extent to which a model can generalize on data it has never seen before. After the model is trained on 80% of the training data, testing on 20% of the test data, which is separate from the training data, helps ensure that the model does not simply memorize the training data (overfitting) but can also make accurate predictions on new, similar data. In this study, the test results are presented in Table 5.

Table 5. Matrix Test Results

Class	Precision	Recall	F1-score	Support
Riau Malay	0,99	0,99	0,99	860
Pontianak Malay	0,99	0,98	0,99	895

The results of the similarity detection between traditional Malay houses in Riau and traditional Malay houses in Pontianak in terms of roof shape, building structure, carvings and ornaments, overall building shape, window shape, and staircase shape are presented in Table 6.

Table 6. Details of Experimental Findings in Identifying the Architectural Style of Traditional Malay Houses

Experiment	Type of Style	Similarity (%)		Split Data
		Riau Malay	Pontianak Malay	
1st	Roof Shape	62,4%	60,5%	80:20
2nd	Building Structure	70.5%	73,3%	80:20
3rd	Carvings and Ornaments	72,3%	68,7%	80:20
4th	Full building of the original	82,3%	80,9%	80:20
5th	window shape	85,3%	83.5%	80:20
6th	staircase shape	97,4%	99,9%	80:20
7th	Full building of the original	85,5%	99,9%	80:20
8th	window shape	99,4%	99,7%	80:20
9th	staircase shape	92,3%	99,4%	80:20
10th	Full building of the original	95,8%	95,8%	80:20
11th	Roof Shape	98,8%	88,7%	70:30
12th	Building Structure	98,9%	98,9%	70:30
13th	Carvings and Ornaments	80,8%	80,8%	70:30
14th	Roof Shape	90,8%	90,8%	70:30
15th	Full building of the original	99,4%	92,4%	70:30
16th	Carvings and Ornaments	99,2%	91,2%	70:30
17th	window shape	99,2%	82,2%	70:30
18th	staircase shape	98,9%	88,9%	70:30
19th	Carvings and Ornaments	99,1%	90,1%	70:30
20th	Full building of the original	99,7%	89,7%	70:30

The test results can be seen in the visualization in Figure 5 below for all experiments that have been conducted with different data splits and object types.

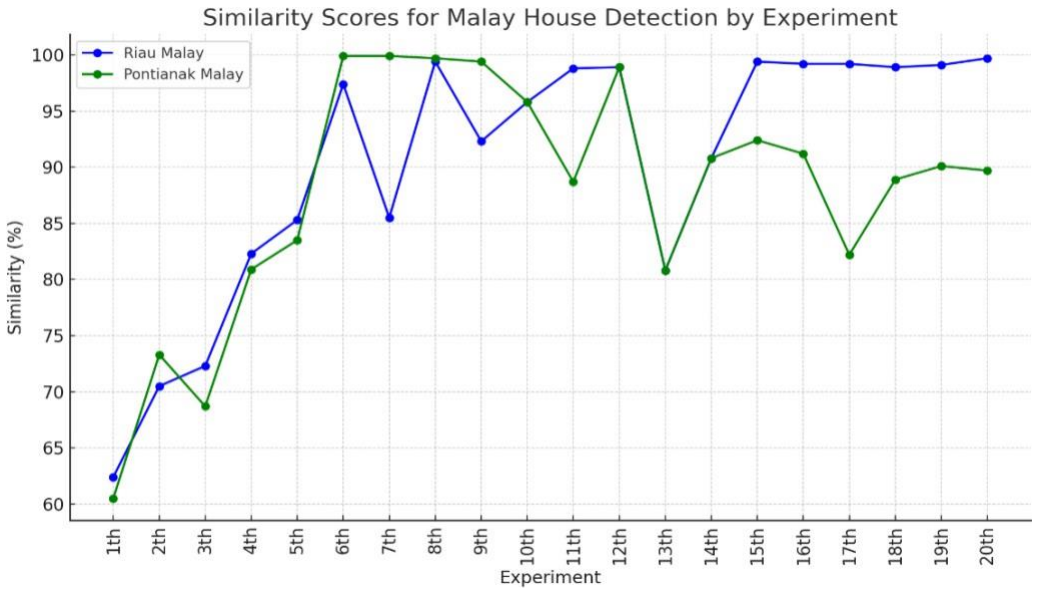

Fig 6. The visualization of the similarity scores for Riau Malay and Pontianak Malay across the 20 experiments

F. Analysis of Model Testing Results

The results of testing the Vision Transformer (ViT) model show excellent performance in identifying architectural styles in traditional Malay houses in Riau and Pontianak. With very high precision, recall, and F1-score (0.99 for both classes), the model shows near-perfect accuracy in recognizing elements of traditional houses. Accuracy reached 99.4%, indicating that the model can correctly predict almost 99.4% of the total data tested. Although there was a slight difference in the amount of test data between the two classes, the model still showed an excellent balance between precision and recall, reflecting its ability to accurately identify positive elements and not miss many important examples.

Overall, the ViT model demonstrates exceptional ability in recognizing and identifying traditional Malay houses in Riau and Pontianak, with highly stable and accurate performance. Overall, the ViT model has successfully identified the architectural styles of traditional Malay houses in Riau and Pontianak with outstanding performance, making it ready for use in real-world applications related to architectural recognition and cultural heritage preservation.

6. Conclusion

This study developed a Vision Transformer (ViT)-based system to identify traditional Malay architectural styles. Using video data of Malay buildings processed through image normalization and augmentation, the deep learning model achieved an accuracy of over 85% or 99.4%. Model evaluation using metrics such as accuracy, precision, recall, and F1-score showed adequate results. However, training with a small amount of data will result in overfitting, which causes objects to be unrecognizable or have low accuracy. After hyperparameter optimization, the model can be integrated into an application for automatic recognition of Malay architectural styles. This research makes a significant contribution to the preservation and digital documentation of traditional buildings, which

can support the tourism and cultural sectors. This research has the potential to be developed in the future by expanding to the identification of other architectural styles, such as predicting the level of damage to buildings based on natural conditions and light direction, and detecting damage to buildings.

Acknowledgment

We would like to thank KemdiktiSainteks for supporting this research in accordance with contract number 114/PDD/LPPM.UAD/VI/2025. We would also like to thank the Research and Community Service Institute of Universitas Ahmad Dahlan for its guidance and support in carrying out this research. We are very grateful to the managers of the traditional houses in Pekanbaru and Pontianak, Riau, for allowing the research team to collect data.

References

[1] S. Alsheikh Mahmoud, H. Bin Hashim, M. F. Shamsudin, and H. Alsheikh Mahmoud, “Effective Preservation of Traditional Malay Houses: A Review of Current Practices and Challenges,” *Sustain*, vol. 16, no. 11, p. 4773, 2024, doi: 10.3390/su16114773.

[2] S. Alsheikh Mahmoud and H. Bin Hashim, “Traditional Malay House Preservation: Guidelines for Structural Evaluation,” *Buildings*, vol. 15, no. 5, 2025, doi: 10.3390/buildings15050782.

[3] Syahputra, Dara Wisdianti, and Cut Nuraini, “Application of Neo-Vernacular Malay Architecture in the Sports Buildings of the Youth and Sports Department of North Sumatra,” *Int. J. Integr. Sci.*, vol. 4, no. 6, pp. 1333–1350, 2025, doi: 10.55927/ijis.v4i6.359.

[4] N. Safiqah binti Abdul Razak and A. Bin Sabil, “Revitalizing Kelantan Malay Traditional Architecture: the Adaptation of Malay Traditional Architecture Design Features in Kelantan to Modern Contemporary Architectural Scheme,” *J. Archit. Plan. Constr. Manag.*, vol. 14, no. 2, 2024, doi: 10.31436/japcm.v14i2.925.

[5] J. Zhong, J. Yin, P. Li, P. Zeng, M. Zang, and R. Luo, “ArchiLense : A Framework for Quantitative Analysis of Architectural Styles Based on Vision Large Language Models,” vol. 1, pp. 1–11, 2025.

[6] E. Kalsum, T. W. Caesariadi, Y. Purnomo, S. A. Gapor, and H. F. A. Rahman, “Regionalism in Architecture: a Study of Local Perceptions on Public State Buildings in West Kalimantan Province, Indonesia,” *Plan. Malaysia*, vol. 22, no. 1, pp. 270–282, 2024, doi: 10.21837/pm.v22i30.1439.

[7] M. B. Starzyńska-Grześ, R. Roussel, S. Jacoby, and A. Asadipour, “Computer Vision-based Analysis of Buildings and Built Environments: A Systematic Review of Current Approaches,” *ACM Comput. Surv.*, vol. 55, no. 13s, 2023, doi: 10.1145/3578552.

[8] S. Ahmad, A. E. Yetti, N. E. Ali, M. S. Sanusi, and P. Y. Samsudin, “Adaptation of traditional Malay house design towards housing comfort and satisfaction,” *Malaysian J. Soc. Sp.*, vol. 18, no. 3, 2022, doi: 10.17576/geo-2022-1803-14.

[9] E. Stylianidis, K. Evangelidis, R. Vital, P. Dafiotis, and S. Sylaiou, “3D Documentation and Visualization of Cultural Heritage Buildings through the Application of Geospatial Technologies,” *Heritage*, vol. 5, no. 4, pp. 2818–2832, 2022, doi: 10.3390/heritage5040146.

[10] T. Penjor, S. Banihashemi, A. Hajirasouli, and H. Golzad, “Heritage building information modeling (HBIM) for heritage conservation: Framework of challenges, gaps, and existing limitations of HBIM,” *Digit. Appl. Archaeol. Cult. Herit.*, vol. 35, no. February, p. e00366, 2024, doi: 10.1016/j.daach.2024.e00366.

[11] H. Li and H. Dong, “Architectural Style Classification Based on Deep Learning,” *Comput. Aided. Des. Appl.*, vol. 22, pp. 16–31, 2025, doi: 10.14733/cadaps.2025.S1.16-31.

[12] K. Al-hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, “Vision transformer architecture and applications in digital health: a tutorial and survey,” *Vis. Comput. Ind. Biomed. Art*, vol. 6, no. 1, 2023, doi: 10.1186/s42492-023-00140-9.

[13] S. Zheng, J. Zhang, R. Zu, and Y. Li, “Vision transformer-enhanced thermal anomaly detection in building facades through fusion of thermal and visible imagery,” *J. Asian Archit. Build. Eng.*, vol. 24, no. 4, pp. 2854–2868, 2025, doi: 10.1080/13467581.2024.2379866.

[14] M. Sun, F. Zhang, F. Duarte, and C. Ratti, “Understanding architecture age and style through deep learning,” *Cities*, vol. 128, p. 103787, 2022, doi: 10.1016/j.cities.2022.103787.

[15] E. Cantemir and O. Kandemir, “Use of artificial neural networks in architecture: determining

the architectural style of a building with a convolutional neural network," *Neural Comput. Appl.*, vol. 36, no. 11, pp. 6195–6207, 2024, doi: 10.1007/s00521-023-09395-y.

[16] H. Xu, H. Sun, L. Wang, X. Yu, and T. Li, "Urban Architectural Style Recognition and Dataset Construction Method under Deep Learning of Street View Images: A Case Study of Wuhan," *ISPRS Int. J. Geo-Information*, vol. 12, no. 7, 2023, doi: 10.3390/ijgi12070264.

[17] S. Wang, J. Zhang, A. N. Tun, and K. Sein, "Research on Identification, Evaluation, and Digitization of Historical Buildings Based on Deep Learning Algorithms: A Case Study of Quanzhou World Cultural Heritage Site," *Buildings*, vol. 15, no. 11, pp. 1–18, 2025, doi: 10.3390/buildings15111843.

[18] F. N. U. Neha and A. Bansal, "Understanding the architecture of vision transformer and its variants: A review," *1st Int. Conf. Innov. Eng. Sci. Technol. Res. ICIESTR 2024 - Proc.*, December 2024, doi: 10.1109/ICIESTR60916.2024.10798341.

[19] S. Wassan, A. Bilal, A. Alzahrani, K. Almohammadi, M. Alrashidi, and S. J. Mousavirad, "A modified vision transformer framework for image-based land cover segmentation in rural architectural design and planning," *Sci. Rep.*, vol. 15, no. 1, pp. 1–21, 2025, doi: 10.1038/s41598-025-19234-w.

[20] B. N. Patro, V. P. Namboodiri, and V. S. Agneeswaran, "SpectFormer: Frequency and Attention is what you need in a Vision Transformer," *Proc. - 2025 IEEE Winter Conf. Appl. Comput. Vision, WACV 2025*, pp. 9543–9554, 2025, doi: 10.1109/WACV61041.2025.00924.

[21] K. Mangalam *et al.*, "Reversible Vision Transformers," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 10820–10830, 2022, doi: 10.1109/CVPR52688.2022.01056.

[22] F. Diker and İ. Erkan, "An Approach With Deep Convolutional Neural Networks for Accurate Architectural Style Classification," *New Des. Ideas*, vol. 8, no. 3, pp. 615–640, 2024, doi: 10.62476/ndi83615.

[23] B. Wang, J. Zhang, R. Zhang, Y. Li, L. Li, and Y. Nakashima, "Improving facade parsing with vision transformers and line integration," *Adv. Eng. Informatics*, vol. 60, no. 999, 2024, doi: 10.1016/j.aei.2024.102463.

[24] S. Mia, "2023 International Conference on Cognitive Computing and Complex Data, ICCD 2023," *2023 Int. Conf. Cogn. Comput. Complex Data, ICCD 2023*, vol. 3, 2023.

[25] A. Wong, S. Abbasi, and S. Nair, "TurboViT: Generating Fast Vision Transformers via Generative Architecture Search," pp. 1–5, 2023, [Online]. Available: http://arxiv.org/abs/2308.11421

[26] O. Chibuike and X. Yang, "Convolutional Neural Network–Vision Transformer Architecture with Gated Control Mechanism and Multi-Scale Fusion for Enhanced Pulmonary Disease Classification," *Diagnostics*, vol. 14, no. 24, pp. 1–17, 2024, doi: 10.3390/diagnostics14242790.

[27] A. Dosovitskiy *et al.*, "An image is worth 16 x 16 words :" *Int. Conf. Learn. Represent.*, pp. 1–21, 2021.

[28] K. Mohiuddin *et al.*, "Retention Is All You Need," in *International Conference on Information and Knowledge Management, Proceedings*, 2023, pp. 4752–4758. doi: 10.1145/3583780.3615497.

[29] B. A. Isnanto and detiksumut, "'Mengenal 5 Rumah Adat Riau: Ciri-ciri Arsitektur dan Filosofinya,'" Detik.com. [Online]. Available: https://www.detik.com/sumut/budaya/d-6841343/mengenal-5-rumah-adat-riau-ciri-ciri-arsitektur-dan-filosofinya

[30] Fajar Pendidikan, "Mengenal Rumah Melayu Atap Lontik Rumah Adat Provinsi Riau," Fajarpendidikan.co.id. [Online]. Available: https://www.fajarpendidikan.co.id/ciri-khas-rumah-melayu-atap-lontik-rumah-adat-provinsi-riau

[31] F. Rahman and H. Kurniawan, "Penerapan Ciri Khas Arsitektur Melayu Pada Fasad Bangunan Kontemporer Di Kota Pekanbaru (Kasus Perkantoran Pemerintahan Di Tenayan Raya)," *J. Archit. Des. Dev.*, vol. 2, no. 2, p. 103, 2021, doi: 10.37253/jad.v2i2.4967.

[32] Jamilah and Nurmaningsih, "Eksplorasi Etnomatematika pada Tenun Corak Insang Melayu Pontianak," *J. Ris. Pembelajaran Mat. Sekol.*, vol. 8, no. 1, pp. 1–9, 2024, doi: 10.21009/jrpms.081.01.

[33] Koransaku, "Menilik Arsitektur Rumah Adat Melayu Kalbar," koransaku.hops.id. [Online]. Available: https://koransaku.hops.id/wisata/pr-3762607172/menilik-arsitektur-rumah-adat-melayu-kalbar?

[34] J. S. Yoo, H. Lee, and S. W. Jung, “Hierarchical Spatiotemporal Transformers for Video Object Segmentation,” *Proc. - 2023 IEEE/CVF Int. Conf. Comput. Vis. Work. ICCVW 2023*, vol. 1, pp. 795–805, 2023, doi: 10.1109/ICCVW60793.2023.00087.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “Detr,” *arXiv*, pp. 1–17, 2020.

[36] S. Aburass, O. Dorgham, J. Al Shaqsi, M. Abu Rumman, and O. Al-Kadi, *Vision Transformers in Medical Imaging: a Comprehensive Review of Advancements and Applications Across Multiple Diseases*, no. 0123456789. Springer International Publishing, 2025. doi: 10.1007/s10278-025-01481-y.

[37] A. Khan *et al.*, “A survey of the vision transformers and their CNN-transformer based variants,” *Artif. Intell. Rev.*, vol. 56, October, pp. 2917–2970, 2023, doi: 10.1007/s10462-023-10595-0.