

Vision Transformer-Based Recognition of Riau Malay Architectural Features with Cross-Regional Comparison for Digital Heritage Documentation

Heri Pramono¹, Sri Winiarti^{2*}, Abdul Fadlil², Sunardi²

¹ Architecture, Star YKPN, Gagak Rimang No 1, Yogyakarta, Indonesia

² Doctoral Program of Informatics, Universitas Ahmad Dahlan, Ringroad Selatan, Tamanan Bantul, Yogyakarta, Indonesia

Article Info

Article history:

Received April 28th, 2026

Revised May 23rd, 2026

Accepted August 02nd, 2026

Keywords:

Architectural Feature Recognition
Digital Heritage Documentation
Pontianak Malay Architecture
Riau Malay Architecture
Vision Transformer

ABSTRACT

Traditional Riau Malay architecture requires systematic digital documentation for heritage preservation. This study evaluates a Vision Transformer (ViT-B/16) model initialised with ImageNet-1K pretrained weights for recognising Riau Malay architectural features, using Pontianak Malay architecture for cross-regional comparison. The dataset was constructed from 24 architectural videos covering roof shapes, building structures, ornaments, windows, staircases, and full-building views. Using automated spatiotemporal segmentation at five frames per second, 13,230 frames were extracted, resized to 224×224 pixels, normalised, augmented, and divided into 16×16-pixel patches. Evaluation on a balanced, held-out test set of 32 clips yielded an overall accuracy of 84.38%, macro precision of 84.51%, macro recall of 84.38%, and macro F1-score of 84.36%. Distinctive elements, such as roofs, windows, staircases, and full buildings, achieved higher recognition performance when clearly visible. Conversely, partially visible structures and detailed ornaments exhibited variable performance due to lighting, viewpoint, and visual complexity. Given the single hold-out split and the limited number of source videos, these findings are preliminary; high feature-specific accuracies should not imply perfect recognition or generalizability. Nonetheless, the results demonstrate ViT-B/16's strong potential to support the digital recognition of Malay architectural heritage. Future work should incorporate grouped five-fold cross-validation, independent building-level testing, CNN baseline comparisons, ROC-AUC analysis, and attention map visualisations.

This is an open-access article under the [Creative Commons Attribution 4.0 International License](#)



Corresponding Author:

Sri Winiarti,
Informatics, Universitas Ahmad Dahlan,
Ringroad Selatan, Yogyakarta, Indonesia
Email: sri.winiarti@uad.ac.id

1. INTRODUCTION

Traditional Malay architecture of Riau is an important part of the cultural heritage, representing visual identity, historical value, and local knowledge, as manifested in the composition of facades, roof geometry, openings, building mass proportions, and various forms of ornamentation. However, the preservation of such architectural knowledge still faces serious challenges, namely the lack of systematic digital documentation, reliance on manual expert identification, and the limited availability of automated

classification tools that can consistently recognise a building's visual features [1][2]. In the field of cultural heritage, classification remains a crucial foundation for archiving, search, HBIM modelling, semantic enrichment, and data-driven conservation; however, manual processes are often slow, subjective, and difficult to scale with increasing numbers of objects. Recent studies also indicate that AI, machine learning, and deep learning are increasingly viewed as strategic tools to support the more efficient and measurable conservation, documentation, and management of cultural heritage [3][4].

In recent years, international research has demonstrated significant progress in the recognition of images of architectural heritage. Siountri and Anagnostopoulos developed a classification system for cultural heritage buildings in Athens using deep learning, demonstrating that the morphological elements of buildings can be automatically learned to distinguish architectural typologies. Luo et al compared several transfer learning methods for cultural heritage image classification and confirmed that the choice of feature extraction strategy is critical to recognition accuracy. Artopoulos et al. also extended AI approaches to classifying built heritage styles in 3D representations, while Sasithradevi et al. introduced MonuNet for classifying Kolkata heritage images, addressing the need for efficient and accurate identification of urban architectural heritage. These studies demonstrate that automatic visual feature extraction has become a primary direction in digital heritage [5][6].

Subsequent developments have moved toward more contextual and nuanced recognition. Ottoni et al. proposed a deep learning approach incorporating transfer learning and data augmentation for classifying cultural heritage buildings in Brazil. They introduced the ImageMG dataset, comprising 6,449 images of 94 historic buildings [7]. Li et al. developed a deep learning method to identify traditional villages across cultural types, while Luo et al. utilised UAV imagery and improved deep learning algorithms to recognise ethnic architectural heritage. Wang et al. employed deep learning to identify, evaluate, and digitise historic buildings at the Quanzhou World Cultural Heritage Site [8, 9, 10]. Han et al. specifically highlighted formal features such as deep eaves, gables, and long windows in Jiangsu vernacular architecture, demonstrating the importance of recognising formal features rather than merely classifying final categories. Han et al. also emphasised the need for chronological classification in traditional vernacular architecture, while Zhou et al. expanded the study to include the recognition of decorative patterns as elements of the visual identity of architectural heritage [11][12][13]. Collectively, these studies underscore that the extraction of architectural visual features is now shifting from the global object level toward more detailed formal attributes and elements.

Nevertheless, most previous studies have focused on cases in Greece, India, Brazil, China, and specific World Heritage sites, while traditional Malay architecture, particularly that of Riau, has received very limited attention in reputable international literature [14][15]. On the other hand, studies on the preservation of traditional Malay houses indicate that these structures face threats from modernisation, urbanisation, material degradation, and weak sustainable conservation strategies, making the need for digital documentation and automated identification increasingly urgent. In other words, there is a real need to bridge Malay architectural conservation studies and developments in computer vision-based AI so that architectural heritage is not only described qualitatively but also mapped, recognised, and classified computationally [16].

From a methodological perspective, most research on heritage image classification from 2023 to 2025 remains dominated by CNNs, transfer learning, object detection, and hybrid models that emphasise local features. In fact, the classification of architectural style and identity very often depends on the spatial relationships between elements distributed across the entire façade, for example, the relationship between roof shapes, the positioning of openings, the rhythm of columns, and ornamentation, which are not always optimally captured using only local convolutional kernels. Recent literature on Vision Transformers confirms that ViTs excel at modelling global dependencies via self-attention mechanisms, while up-to-date surveys also indicate that both ViT variants and CNN–ViT hybrids tend to strike a good balance among accuracy, data efficiency, and the ability to capture broader visual context. In the domain of cultural heritage, Khan et al. have demonstrated through ViT-HVE that Vision Transformers can be used not only for recognition but also for evaluating heritage value; Fan et al. have also shown that transformer-based multimodal approaches open up opportunities for richer cultural heritage recognition [17]. This provides a strong methodological foundation for applying Vision Transformers to the recognition of features of traditional Riau Malay architecture [18].

Based on this review, the research gaps can be summarised as follows. First, there remains a scarcity of international research that specifically targets traditional Riau Malay architecture as the subject of AI-based image classification. Second, previous studies have largely focused on the general classification of

buildings or sites, but have not sufficiently emphasised the extraction of distinctive architectural features that explain why a building is recognised as Riau Malay. Third, the application of Vision Transformers in the domain of regional traditional architecture remains limited, even though these models are theoretically better suited to capturing global relationships among visual components. Fourth, the integration of automatic classification with digital architectural initiatives, such as digital documentation, indexing, and HBIM support, has not yet been widely implemented in the context of Malay architectural heritage [8][19].

Based on this gap, the novelty of this research lies in the development of an automated feature extraction and classification approach to recognise traditional Riau Malay architecture using Vision Transformers, with a focus not only on class prediction, but also on the model's ability to capture the distinctive visual characteristics of buildings in a more representative manner on a global scale [16][20]. Another novelty of this research is the application of the transformer paradigm to Riau Malay architectural heritage objects that are still rarely studied in international publications, as well as the placement of the classification results as a basis for digital architectural documentation and data-based conservation. Thus, this research not only adopts the latest technology but also fills gaps in the study object, feature level, and cultural-area context [21].

The objectives of this research are to develop a Vision Transformer-based model to extract visual features and automatically classify traditional Riau Malay buildings; evaluate the model's ability to distinguish visual characteristics among buildings; and develop a computational framework to support the documentation and preservation of digital architecture. The benefits of this research include: (1) academic benefits, namely enriching the literature on the intersection of AI, computer vision, and digital architecture; (2) practical benefits, namely providing an assistive approach for the inventory, archiving, and rapid identification of traditional buildings; and (3) cultural benefits, namely supporting efforts to preserve the identity of Riau Malay architecture through a more systematic, scalable, and replicable digital approach. The importance of this research is further underscored by recent studies confirming that the annotation, classification, and automatic recognition of cultural artefacts are crucial for information retrieval, scene understanding, public exploration, and the development of intelligent systems in the cultural heritage ecosystem [22][23].

Ottoni et al., Han et al., and Zhou et al. demonstrate the importance of focusing on architectural elements and formal features as markers of visual identity [24][6]. Shehata et al. and Sasithra et al. emphasise the relevance of classification for the heritage digitisation ecosystem and HBIM. Wang and Fan demonstrate that cultural heritage research is now moving towards more intelligent, evaluative, and multimodal models. Meanwhile, the surveys of Li and Dong and Montrezol and Oliveira provide methodological justification for why Vision Transformers are suitable for image classification tasks that require an understanding of global context. This study then integrates all these contributions into a more specific context, namely, traditional Riau Malay architecture [17][15].

Thus, this research is important because it addresses three needs simultaneously: the scientific need to expand the application of Vision Transformers to the classification of traditional architecture, the technological need to provide a more automated and scalable building characteristic identification system, and the cultural need to maintain the sustainability of Riau Malay architectural knowledge amidst the pressures of changing times. This research is expected to serve as a bridge between cultural heritage preservation and artificial intelligence, so that traditional architecture is not merely a passive object of documentation but becomes visual knowledge that can be analysed, archived, and reused digitally [13][1].

2. RESEARCH METHOD

2.1. Research Design

This study was designed as an exploratory architectural feature recognition study rather than a complete digital heritage preservation system. The primary objective was to examine whether a Vision Transformer could generate distinguishable representations of Riau Malay architectural features. Pontianak Malay architectural images were included as a cross-regional comparison domain to examine variations in visual recognition scores. The model's outputs were subsequently discussed as potential metadata for digital heritage documentation; however, HBIM implementation and automated conservation decision-making were outside the experimental scope of this study.

This study develops an AI-based digital architectural approach to identify and classify distinctive features of traditional Riau Malay buildings using a Vision Transformer. Video recordings are converted into

Vision Transformer-Based Recognition of Riau Malay Architectural Features with Cross-Regional Comparison for Digital Heritage Documentation (Heri Pramono)

image frames to capture variations in viewpoint, lighting, and environmental context. Each frame is divided into non-overlapping $16 \times 16 \times 16$ -pixel patches, which are projected into embedding vectors and combined with positional encoding to preserve spatial relationships among architectural elements. A classification token is added to aggregate global image information. At the same time, multiple Transformer encoder layers employ multi-head self-attention, residual connections, normalisation, and feed-forward networks to model relationships among roof geometry, façade composition, openings, structural elements, and ornaments. The final representation of the classification token is then used to determine the architectural class. The complete research workflow is presented in Figure 1.

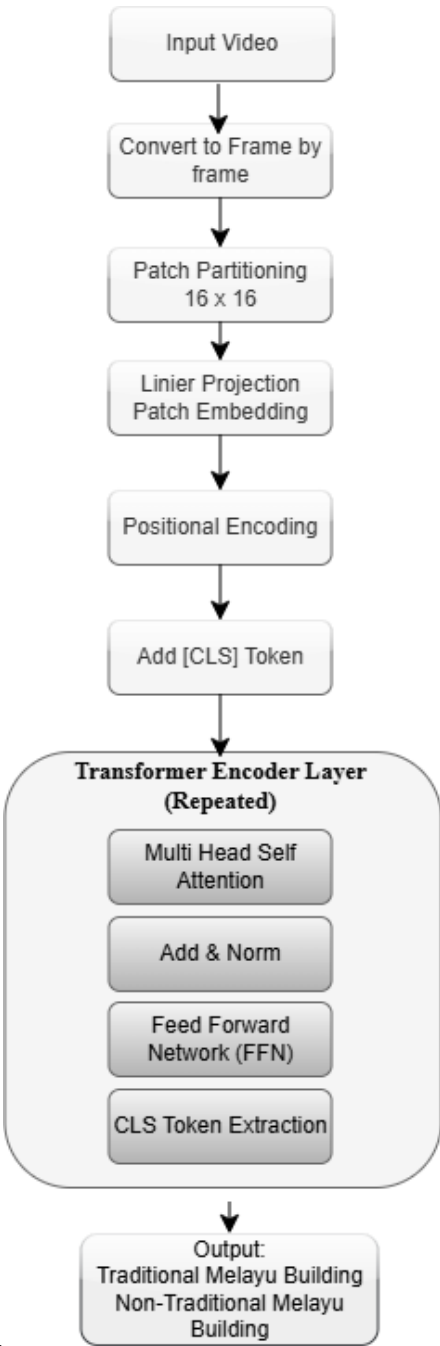


Figure 1. Research Flow: Recognising the Architectural Style of Riau and Pontianak Malay Traditional Buildings with Vision Transformer

The model groups buildings into two categories: Traditional Riau Malay and Traditional Pontianak Malay. More than just a classification, this approach enables the identification of visual characteristics that differentiate the two categories, thereby supporting the digital architectural interpretation process. Thus, this method serves not only as a classification tool but also as a means of understanding the visual patterns and structures that shape the identity of Riau Malay architecture.

2.2. Dataset Construction and Data Sources

The dataset used in this study was constructed from video documentation of traditional Malay buildings representing Riau Malay and Pontianak Malay architecture, serving as comparative data from Kalimantan. The recordings focused on architectural components considered relevant for distinguishing Malay architectural characteristics, including roof shape, building structure, carvings and ornaments, window shape, staircase shape, and full-building views. The duration of the videos varied according to the size, visual complexity, and number of viewpoints required to document each architectural feature.

Dataset construction was performed automatically using a spatiotemporal processing approach, rather than through manual frame selection. The procedure analysed both the temporal continuity of consecutive video frames and the spatial changes occurring within the recorded scenes. Temporal information was used to identify visually coherent sequences and transitions between camera shots. In contrast, spatial information was used to distinguish changes in viewpoint and architectural regions, such as front views, side views, roofs, windows, staircases, and ornament details. Consequently, video segmentation and frame extraction were performed systematically based on variations in time, camera movement, viewpoint, and the building's visual composition. The process generated a total of 13,230 image frames, consisting of 1,890 roof-shape frames, 2,050 building-structure frames, 850 carving-and-ornament frames, 3,625 full-building frames, 3,270 window-shape frames, and 1,545 staircase-shape frames. The use of an automated extraction procedure reduced dependence on subjective manual selection and improved the consistency and reproducibility of dataset construction. The number of frames corresponds to the video durations and the fixed extraction rate presented in Table 1.

The spatiotemporal approach grouped consecutive frames with similar viewpoints and spatial composition into visually homogeneous segments. Significant changes in camera angle or architectural content defined new segment boundaries. Each segment represented a consistent view, such as a façade, roof, ornament, window, staircase, or side view. This automated process was applied before the ViT input to reduce variations caused by abrupt camera movements and scene changes. Following extraction, segmentation, and quality filtering, 160 eligible video clips were retained for model development and evaluation. The dataset was initially divided using a grouped hold-out strategy, with 80% of the clips assigned to the development set and 20% assigned to the held-out test set. This procedure yielded 128 development clips and 32 test clips. The development set was subsequently divided into 112 training clips and 16 validation clips. Therefore, the final data distribution comprised 112 training clips (70%), 16 validation clips (10%), and 32 test clips (20%). The held-out test set was balanced, comprising 16 Riau Malay clips and 16 Pontianak Malay clips. All clips originating from the same building, source video, or spatiotemporal segment were assigned to the same subset to reduce the risk of data leakage. All clips originating from the same source building, source video, or spatiotemporal segment were retained in the same subset to reduce the risk of data leakage, as shown in Table 1.

Table 1. Distribution of training, validation, and testing data

Subset	Number of clips	Overall percentage
Training	112	70%
Validation	16	10%
Testing	32	20%
Total	160	100%

Based on Table 1, the dataset comprised 160 eligible video clips, which were divided into training, validation, and test subsets. A total of 112 clips, representing 70% of the dataset, were allocated to the training subset and used to optimise the model parameters. The validation subset contained 16 clips, equivalent to 10% of the total dataset, and was used to monitor model performance, select the best checkpoint, and mitigate overfitting during training. The remaining 32 clips, representing 20% of the dataset,

were assigned to the held-out testing subset and used exclusively for the model's final evaluation. Overall, the distribution ensured that most of the data were available for model learning while separate validation and testing subsets were retained for model selection and unbiased performance assessment. Under the grouped hold-out strategy, all clips originating from the same building, source video, or spatiotemporal segment were maintained within the same subset to reduce the risk of data leakage.

2.3. Shot and Angle Segmentation

Shot- and angle-based segmentation was performed automatically using a two-level spatiotemporal pipeline. Each source video was divided into five-second clips and sampled at 5 FPS, producing up to 25 frames per clip. Sixteen frames were then uniformly selected to standardise the temporal input and reduce computational requirements. The selected frames were grouped into visually homogeneous segments based on temporal continuity, camera viewpoint, and architectural content, including front, side, roof, full-building, partially visible, and ornament-focused views.

Relevant architectural regions were defined as regions of interest covering façades, roofs, windows, staircases, structural elements, and ornaments. After quality checking, the frames were converted to RGB format, resized to 224×224 pixels, and separated when abrupt transitions or inconsistent viewpoints were detected. Gaussian blur with a 7×7 kernel and $\sigma=1.5$ was applied outside the main architectural region to reduce background interference.

Each sequence was stored with its source video, regional domain, architectural feature, viewpoint, and segment identifiers. Frames from the same spatiotemporal segment were kept within a single subset for training, validation, or testing to reduce data leakage. This rule-based pipeline improved the consistency and reproducibility of dataset construction while reducing reliance on manual frame selection.

2.4. Image Preprocessing

Each video frame was converted to RGB format and resized to 224×224 pixels. The resized image was then divided into non-overlapping patches of size 16×16 pixels. Because each image had dimensions of 224×224 pixels, the patching process produced 14×14 (196) patches per image. Each patch contained three colour channels, namely red, green, and blue (RGB). Therefore, a single patch consisted of $16 \times 16 \times 3 = 768$ pixel values. The pixel values within each patch were flattened into a 768-dimensional input vector. This vector was subsequently linearly projected into a 768-dimensional embedding representation before being processed by the Transformer encoder [25][26].

Data augmentation was applied only to the training subset and included horizontal flipping, rotation, cropping, colour adjustment, and Gaussian blur. Gaussian blur was applied using a 7×7 kernel, $\sigma=1.5$, and an application probability of 0.5. Resizing, RGB conversion, patch generation, and normalisation were consistently applied to the training, validation, and testing subsets to ensure uniform input dimensions and data distributions. The main input specifications were as follows: a patch size of 16×16 pixels, three RGB channels, 196 patches per image, 768 input values per patch, and a 768-dimensional patch embedding.

ViT-B/16 comprises 12 Transformer encoder blocks, 12 self-attention heads, a hidden dimension of 768, and an MLP dimension of 3,072. The model was trained using weighted cross-entropy loss. During the frozen-backbone stage, the classifier was optimised using Adam with a learning rate of 3×10^{-4} . During fine-tuning, AdamW was used with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4} . The checkpoint with the highest validation F1-score was retained.

2.5. Vision Transformer Architecture

The model used in this study is the Vision Transformer Base with a patch size of 16×16 pixels (ViT-B/16). This architecture was implemented using `models.vit_b_16(weights="IMAGENET1K_V1")`, so the model's initial weights were derived from the pre-training process on the ImageNet-1K dataset. The use of pre-trained weights allows the model to leverage initial visual knowledge, such as the shape, edges, texture, colour, and structure of objects, before being fine-tuned on a dataset of traditional Malay buildings. Thus, the model is not trained from random weights from the start.

The model input consisted of video clips represented as a five-dimensional tensor with the form (B, T, C, H, W), where B denotes the batch size, T denotes the number of frames per clip, C represents the number of colour channels, and H and W represent the frame height and width. In this study, B=4, T=16, C=3, and H = W = 224. Therefore, the input tensor had the dimensions (4, 16, 3, 224, 224).

Before being processed by the ViT backbone, the batch dimension and temporal dimension are combined. The input tensor is transformed from (4, T, 3, 224, 224) to (4T, 3, 224, 224). This reshaping allows each frame to be processed individually by the ViT backbone while maintaining its association with the original video clip. Each clip supplied to the model consisted of 16 uniformly sampled frames. Therefore, with a batch size of 4, each training iteration processed $4 \times 16 = 644$ frames through the ViT backbone. Accordingly, the input tensor had the dimensions (4, 16, 3, 224, 224) before the batch and temporal dimensions were combined. Each frame is first converted to RGB format and resized to 224×224 pixels. Next, the frame is divided into non-overlapping image patches, each measuring 16×16 pixels. Thus, each frame yields 196 patches, with each patch containing three RGB colour channels and therefore 768 pixel values.

These values are flattened into a one-dimensional vector, then transformed through a linear projection layer into a 768-dimensional embedding vector. Although the number of input values per patch and the embedding dimension are both 768, the linear projection process still converts the raw pixel values into a feature representation learned by the model. Thus, a single frame is represented as a sequence of 196 patch tokens, each with 768 feature values. ViT then adds a special token called the classification token or [CLS] token at the beginning of the patch sequence. This addition brings the total number of tokens processed for each frame to 197 tokens: 196 patch tokens and one [CLS] token. The [CLS] token gathers information from all parts of the image and, at the end of the process, serves as a global representation of the entire image. In addition, positional embeddings are added to each token so that the model can retain information about the spatial position of each patch. This positional information helps the model distinguish the top, bottom, left, right, and centre of the building image.

The token sequence is then processed through Transformer encoder blocks consisting of a multi-head self-attention mechanism, layer normalisation, residual connections, and a multilayer perceptron. The self-attention mechanism allows each patch to interact with all other patches within a single frame. Consequently, the model can learn global relationships among architectural elements, such as roof shapes, window positions, column arrangements, ornamentation, staircases, and the overall composition of the building's facade. Mathematically, the self-attention mechanism can be expressed as [27]:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^t}{\sqrt{d_k}}\right)V \quad (1)$$

In the attention mechanism, Query (Q) represents the feature vectors used to retrieve relevant information from the input sequence, while Key (K) serves as the reference vectors against which the Query is compared to determine the relevance of each input element. Value (V) contains the feature representations that are propagated to the output according to the computed attention weights. The Key transpose K^t is multiplied by the Query matrix to produce the dot-product QK^t , which measures the similarity between the corresponding feature representations. The resulting scores are then scaled by the square root of the Key dimension d_k to prevent excessively large values and to stabilise the optimisation process. Subsequently, the Softmax function normalises the scaled scores into a probability distribution, producing attention weights that indicate the relative importance of each feature. Finally, the Attention (Q, K, V) output is obtained by aggregating the Value vectors weighted by these attention scores, yielding a contextualised feature representation that emphasises the most relevant information for the classification task. The original ImageNet classification heads are removed using:

```
self.backbone.heads=nn.Identity().
```

Therefore, the ViT backbone no longer produces predictions for the 1,000 ImageNet classes but instead generates a 768-dimensional feature vector for each frame. The output shape of the backbone at this stage is (B × T, 768). This feature vector is then converted back to the video sequence format from (B × T, 768) to (B, T, 768). The 768-dimensional video representation is then fed into a fully connected classification layer defined as `nn.Linear(768, num_classes)`. This layer produces logit values for each target class, the types of traditional Malay buildings defined in this study. The class with the highest logit value is selected as the final prediction. Model training was conducted in two stages over a total of ten epochs. During the first five epochs, the ViT backbone weights were frozen, allowing the learning process to focus on adjusting the classification layer. From the sixth to the tenth epoch, the backbone was reactivated for fine-tuning, thereby adapting the pre-trained ImageNet features to the visual characteristics of traditional Malay buildings. A batch size of 4 was chosen so that each iteration could process multiple video clips simultaneously while still accounting for the computational memory requirements of ViT-B/16 and the number of frames in each clip.

The training results in the report show a decrease in training loss after the backbone was reactivated during fine-tuning. Overall, this method advances digital architecture by enabling data-driven building analysis, automating the recognition of distinctive Malay architectural features, and generating metadata and classification labels to support digital heritage documentation and preservation.

Figure 2 illustrates an integrated digital architectural framework for the preservation of cultural heritage buildings, combining digital data acquisition technologies, Vision Transformer (ViT)-based analysis, semantic classification of architectural elements, and integration into a Heritage Building Information Modelling (HBIM) system. This workflow demonstrates how visual data from building objects is progressively processed into structured information that not only supports the automatic identification of architectural patterns but also enables sustainable digital storage, analysis, and preservation.

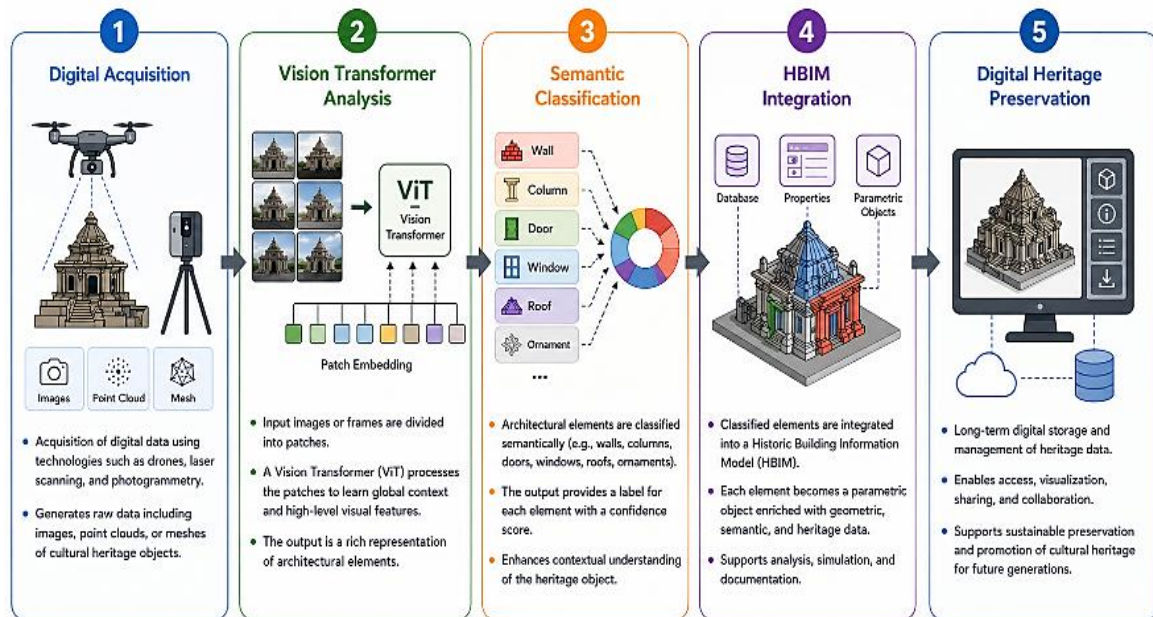


Figure 2. A Digital Architecture Model Integrating Heritage Building Information Modelling (HBIM) with the Vision Transformer Method

2.6. Model Evaluation

The model was evaluated using a grouped hold-out procedure. The dataset was initially divided into an 80% development set and a 20% held-out test set. The development set was further separated into training and validation subsets for model optimisation and checkpoint selection. The final evaluation was conducted exclusively on 32 held-out clips, comprising 16 Riau Malay clips and 16 Pontianak Malay clips. Accuracy, precision, recall, F1-score, and the confusion matrix were calculated from predictions on this held-out test set.

3. RESULTS AND DISCUSSION

3.1. Data Collection Results

The results of this study indicate that the Vision Transformer (ViT)-based approach holds significant potential to support the transformation of architectural analysis toward a data-driven paradigm. In this context, architecture is no longer understood solely through visual descriptions and subjective interpretations, but also through computational representations that systematically capture building morphology. The automatic feature extraction process performed by ViT enables the identification of architectural elements such as roof geometry, facade composition, and ornament distribution, which are traditionally analysed through manual observation by architectural experts. The collected video data is converted into frames using building features, namely roof shape, ornamentation, staircase shape, and

window shape. Table 2 shows the data acquisition results from traditional Malay buildings, while Table 2 shows the data segmentation results from various views. Table 2 summarises the characteristics of the video dataset used in this study. The dataset comprises 24 videos representing two traditional Malay architectural domains, namely Riau Malay (Video IDs 001–012) and Pontianak Malay (Video IDs 013–024). Each video was recorded at a sampling rate of 5 frames per second (FPS), with durations ranging from 34 to 150 seconds, resulting in a total of 13,230 extracted frames.

Table 2. Data on Traditional Malay Houses of Riau and Pontianak

Video ID	Region/domain	Architectural feature	Duration	Extracted frames	Region of interest
001	Riau Malay	Roof Shape	120 seconds x 5 FPS	600 frames	Front view
002	Riau Malay	Building Structure	140 seconds x 5 FPS	700 frames	Front view
003	Riau Malay	Carvings and Ornaments	85 seconds x 5 FPS	425 frames	Ornamental patterns
004	Riau Malay	Full building of the original	90 seconds x 5 FPS	450 frames	Front view
005	Riau Malay	Window shape	127 seconds x 5 FPS	635 frames	Front view
006	Riau Malay	Staircase shape	34 seconds x 5 FPS	170 frames	Front view
007	Riau Malay	Full building of the original	110 seconds x 5 FPS	550 frames	Side view
008	Riau Malay	Window shape	130 seconds x 5 FPS	650 frames	Side view
009	Riau Malay	Staircase shape	75 seconds x 5 FPS	375 frames	Top view
010	Riau Malay	Full building of the original	150 seconds x 5 FPS	750 frames	Side view
011	Riau Malay	Roof Shape	108 seconds x 5 FPS	540 frames	Top view
012	Riau Malay	Building Structure	140 seconds x 5 FPS	700 frames	Full view
013	Pontianak Malay	Roof Shape	65 seconds x 5 FPS	325 frames	Partially visible
014	Pontianak Malay	Building Structure	130 seconds x 5 FPS	650 frames	Full view
015	Pontianak Malay	Carvings and Ornaments	85 seconds x 5 FPS	425 frames	Ornamental patterns
016	Pontianak Malay	Full building of the original	115 seconds x 5 FPS	575 frames	Full view
017	Pontianak Malay	Window shape	122 seconds x 5 FPS	610 frames	Full view
018	Pontianak Malay	Staircase shape	90 seconds x 5 FPS	450 frames	Side view
019	Pontianak Malay	Full building of the original	110 seconds x 5 FPS	550 frames	Ornamental patterns
020	Pontianak Malay	Window shape	150 seconds x 5 FPS	750 frames	Side view
021	Pontianak Malay	Staircase shape	110 seconds x 5 FPS	550 frames	Side view
022	Pontianak Malay	Full building of the original	150 seconds x 5 FPS	750 frames	Front view
023	Pontianak Malay	Roof Shape	85 seconds x 5 FPS	425 frames	Ornamental patterns
024	Pontianak Malay	Window shape	125 seconds x 5 FPS	625 frames	Front view
Total				13.230	

The dataset includes six architectural feature categories: roof shape, building structure, carvings and ornaments, full building, window shape, and staircase shape. To improve feature diversity, representative frames were selected from multiple viewpoints, including front, side, top, full, partially visible, and ornamental patterns. The number of extracted frames per video ranges from 170 to 750, depending on the recording duration. This combination of architectural features and viewpoints provided visual variation for model development and evaluation within the current dataset. Overall, the table shows that the dataset exhibits high diversity in video duration, number of frames, architectural features, and camera angles. This diversity broadens the range of architectural viewpoints represented in the dataset; however, it does not by itself eliminate the possibility of overfitting or establish generalisation to independent buildings. Table 3 shows the segmentation results using shot and angle.

Table 3. Shot-based Segmentation Results for Traditional Malay Buildings

Label	Results of Visual Segmentation for Traditional Pontianak Malay	Results of Visual Segmentation for Traditional Riau Malay
(1) Side view		
(2) Front view of the building		
(3) Roof		
(4) Ornament		

3.2. Video Data Processing into Frames

Videos were segmented according to camera shots and viewing angles, such as front, left, and right views, to produce visually homogeneous sequences and improve analytical consistency. Before being supplied to the ViT model, each source video was divided into five-second clips, and frames were extracted at 5 FPS. Consecutive frames were then grouped according to temporal continuity and the dominant camera viewpoint. Relevant architectural regions of interest were identified, and Gaussian background suppression was applied outside these regions to reduce visual interference. This procedure-maintained consistency within each frame sequence and supported more effective temporal feature processing. The complete classification workflow for Riau and Pontianak Malay traditional buildings using the Vision Transformer is illustrated in Figure 3.

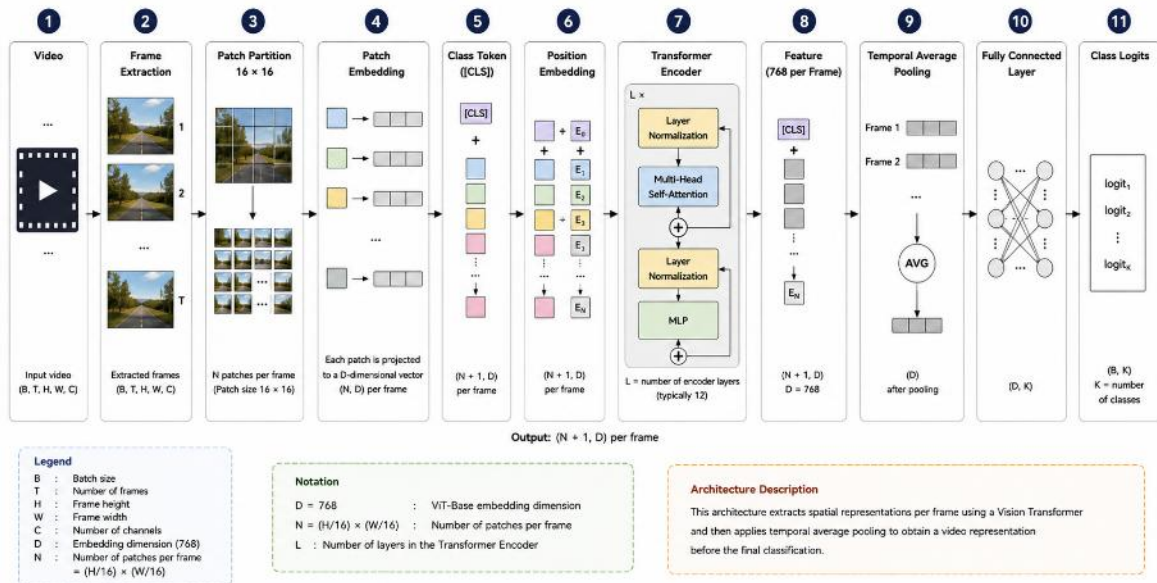


Figure 3. Video Data Processing Workflow for Traditional Malay Buildings

Based on Figure 3, the implementation in Python is shown in Program Code 1.

```
class ViTTemporal(nn.Module):
    def __init__(self, num_classes):
        super().__init__()
        self.backbone = models.vit_b_16(weights="IMAGENET1K_V1")
        self.backbone.heads = nn.Identity()
        self.classifier = nn.Linear(768, num_classes)
    def forward(self, x):
        # x: (B, T, C, H, W)
        B, T, C, H, W = x.shape
        x = x.view(B * T, C, H, W)
        features = self.backbone(x) # (B*T, 768)
        features = features.view(B, T, -1)
        features = features.mean(dim=1) # temporal pooling
        return self.classifier(features)
```

Program Code 1. Classification of Traditional Malay Buildings Using ViT Architecture

Program Code 1 implements a ViT-Temporal model based on a pretrained ViT-B/16 with ImageNet1K_V1 weights. The original classification head is replaced with an identity layer and a new linear classifier. Video inputs in (B, T, C, H, W) format are reshaped so that each frame is processed independently by the ViT backbone, producing a 768-dimensional feature vector per frame. These features are then reorganised into temporal sequences and averaged across frames using temporal mean pooling. The resulting video-level representation is passed to the classifier to predict the traditional Malay building class. Before classification, pre-processing is applied to reduce background interference, suppress noise, and emphasise relevant architectural features. Figure 4 illustrates the data processing steps applied, starting with the image extracted from the video (original frame), through shot-angle segmentation, to the application of Gaussian Blur to the segmentation results.

In the first stage, the original frame is an image extracted from a video of a traditional Malay building, retaining all visual information, including the main object, vegetation, sky, and surrounding environment. At this stage, the image has not undergone object segmentation and therefore still contains elements that could introduce noise during model training.



Figure 4. Segmentation Results Using Shot and Angle Segmentation and Gaussian Blur on Traditional Malay Buildings in Riau

The building is identified as the region of interest to preserve relevant architectural features, including roofs, structures, windows, staircases, and ornaments. Gaussian blur with a 7×7 kernel and $\sigma=1.5$ is applied outside the main object to reduce background interference. This process standardises the visual input and improves feature consistency within the current dataset, although its effect on external generalisation requires further evaluation.

3.3. Classification Performance and Discussion

The dataset comprised six architectural feature categories: roof shape, building structure, carvings and ornaments, full-building view, window shape, and staircase shape. These features were recorded from several viewpoints, including front, side, top, full, partially visible, and ornament-focused views. The dataset composition provided visual variation for model development, although the distribution across feature categories was unequal. Full-building and window representations were more frequently available, whereas only two source videos represented carvings and ornaments. Therefore, the interpretation of feature-related performance should consider both the visibility of the architectural elements and the number of available source videos.

The final classification performance was evaluated on a balanced held-out test set comprising 32 clips: 16 Riau Malay and 16 Pontianak Malay clips. The model correctly classified 14 Riau Malay clips and 13 Pontianak Malay clips. Two Riau Malay clips were incorrectly predicted as Pontianak Malay, while three Pontianak Malay clips were incorrectly predicted as Riau Malay. These results produced an overall accuracy of 84.38%, a macro precision of 84.51%, a macro recall of 84.38%, and a macro F1-score of 84.36%. For the Riau Malay class, the model achieved a precision of 0.8235, a recall of 0.8750, and an F1-score of 0.8485. The recall value indicates that most of the actual Riau Malay clips were successfully identified. However, the lower precision indicates that several Pontianak Malay clips shared visual characteristics that led to their classification as Riau Malay. For the Pontianak Malay class, the model achieved a precision of 0.8667, a recall of 0.8125, and an F1-score of 0.8387. The relatively high precision indicates that predictions assigned to the Pontianak Malay class were generally reliable. In contrast, the lower recall shows that several actual Pontianak Malay clips were not successfully recognised.

The relatively similar F1-scores for both classes indicate that the model produced reasonably balanced classification performance across the two regional architectural domains. Nevertheless, the differences in recall suggest that several visual characteristics were more consistently recognised in the Riau Malay clips than in the Pontianak Malay clips within the current held-out test set. The classification errors may be associated with similarities in roof geometry, façade composition, window configuration, staircase structure, and decorative elements shared across Malay architectural traditions.

Architectural feature visibility also influenced the distinctiveness of the visual representations. Complete and unobstructed views provided a broader context regarding roof form, façade composition, openings, structural proportions, and staircase placement. In contrast, partially visible structures and ornament-focused views provided a more limited global context. They were more sensitive to camera distance, lighting, image resolution, background complexity, and variations in motif scale.

The carvings and ornaments category represented a particularly challenging fine-grained recognition problem because decorative elements contain small motifs, repetitive patterns, subtle geometric differences, and detailed textures. This interpretation is consistent with Zhou et al. [12], who examined the recognition

of decorative patterns in historical buildings in Macau using 11,807 images and several deep-learning architectures, including ResNet50, EfficientNet-B7, ViT-B/16, Swin-B, RegNet-Y-16GF, and ConvNeXt variants. Their findings showed that ConvNeXt-L was effective in recognising complex decorative patterns because it preserves detailed local features while incorporating broader contextual information.

In the present study, the ViT-B/16 modelled long-range relationships among image patches through its self-attention mechanism. However, recent studies indicate that although Vision Transformers are effective at capturing global contextual information, they may be less effective at preserving fine local details when trained on limited datasets [28, 29]. This limitation is particularly relevant to the recognition of carvings and ornaments, which commonly contain small motifs, repetitive structures, subtle geometric variations, and fine-grained textures.

Only two video sources represented the carvings and ornaments category in the present dataset. Therefore, difficulties in recognising these features should not be attributed solely to the ViT-B/16 architecture. Zhou et al. [12] demonstrated that fine-grained architectural ornament recognition benefits from a substantially larger ornament-focused dataset and from architectures capable of integrating local texture extraction with broader contextual information. In addition, variations in camera viewpoint, illumination, partial visibility, and recording conditions can affect the robustness of local visual feature representation [35]. The limited number of independent ornamental observations, variations in motif scale, and differences in acquisition conditions may therefore have contributed to the observed classification errors. A larger and more diverse ornament-focused dataset would provide a stronger empirical basis for evaluating the model's ability to distinguish fine decorative characteristics.

This study did not include a CNN baseline under the same experimental protocol; therefore, the results do not demonstrate the superiority of ViT-B/16 over convolutional models. Differences in performance between Riau Malay and Pontianak Malay performance may reflect architectural characteristics, acquisition conditions, dataset composition, or model behaviour [28, 29]. Future studies should compare ViT-B/16 with CNN and hybrid models using identical data splits, preprocessing, augmentation, optimisation, and evaluation procedures. The videos were also categorised by their dominant architectural features, and their distribution is presented in Table 4.

Table 4. Distribution of Architectural Features

Architectural feature	Number of Videos	Percentage
Roof Shape	4	16.7%
Building Structure	3	12.5%
Carvings and Ornaments	2	8.3%
Full Building	6	25.0%
Window Shape	5	20.8%
Staircase Shape	4	16.7%
Total	24	100%

Table 4 presents 24 videos across six architectural feature categories. Full Building has the largest representation with six videos, followed by Window Shape with five, Roof Shape and Staircase Shape with four each, Building Structure with three, and Carvings and Ornaments with two. The dataset includes both global and local architectural features, although their distributions are unequal. Therefore, findings related to carvings and ornaments should be interpreted cautiously. Figure 5 visualises this distribution and provides an overview of the dataset composition used for model training and evaluation.

Overall, the distribution demonstrates that the dataset encompasses both global architectural features, such as the complete building appearance, and local features, including roofs, windows, staircases, structural elements, and ornaments. Although sample sizes are not perfectly balanced across categories, all key architectural characteristics of traditional Malay buildings are represented. This diversity enables the proposed deep learning model to learn complementary visual features, thereby improving its robustness in recognising buildings under different viewpoints and environmental conditions.

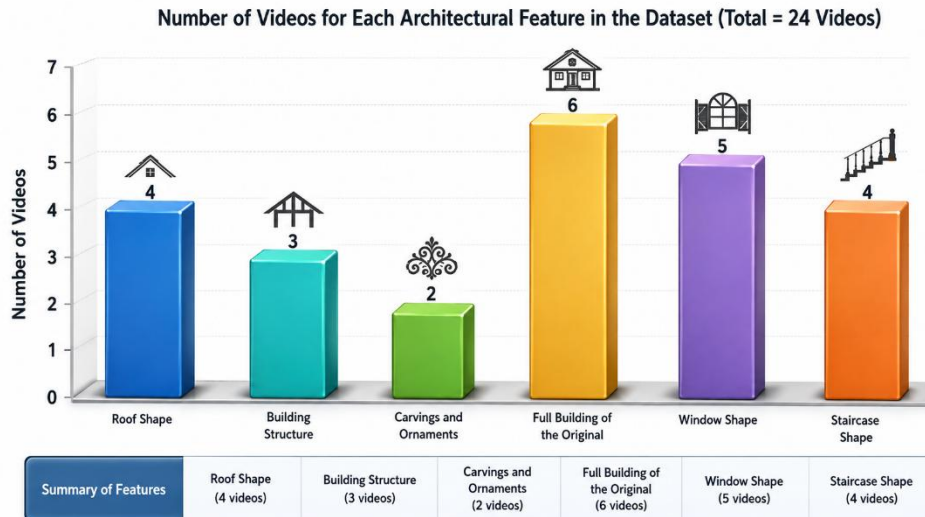


Figure 5. Distribution Chart of Feature Architectures

In this study, the Scikit-learn library was used to calculate the confusion matrix, classification report, and accuracy score, thereby providing information on Precision, Recall, F1-score, and Accuracy as indicators of model performance. The implementation of this evaluation process is shown in Program Code 2.

```
from sklearn.metrics import (
    confusion_matrix,
    classification_report,
    accuracy_score
)

cm = confusion_matrix(y_true, y_pred)
report = classification_report(
    y_true,
    y_pred,
    target_names=["Riau Malay", "Pontianak Malay"]
)
accuracy = accuracy_score(y_true, y_pred)
```

Based on Program Code 2, the evaluation process uses the `confusion_matrix`, `classification_report`, and `accuracy_score` functions from `sklearn.metrics` library. The `confusion_matrix()` function summarises the model's classification results by comparing the true labels (`y_true`) with the predicted labels (`y_pred`). Subsequently, `classification_report()` computes the Precision, Recall, and F1-score for the Riau Malay and Pontianak Malay classes, while `accuracy_score()` calculates the overall classification accuracy. Together, these metrics provide a comprehensive evaluation of the model's classification performance.

Table 5. Confusion Matrix Evaluation Results

Class	TP	FP	FN	TN	Precision	Recall	F1-score	Support
Riau Malay	14	3	2	13	0.8235	0.8750	0.8485	16
Pontianak Malay	13	2	3	14	0.8667	0.8125	0.8387	16
Macro average	—	—	—	—	0.8451	0.8438	0.8436	32
Weighted average	—	—	—	—	0.8451	0.8438	0.8436	32

To evaluate the model's ability to distinguish between two classes of traditional Malay buildings, this study used a confusion matrix to calculate evaluation metrics. In addition to showing the counts of correct and incorrect predictions, the confusion matrix was used to calculate Precision, Recall, F1-score, and Support, thereby providing a more comprehensive picture of the model's classification performance. The evaluation results are presented in Table 5.

Based on Table 5, the model showed relatively balanced performance across both classes. For Riau Malay, precision, recall, and F1-score were 0.8235, 0.8750, and 0.8485, respectively, with 14 of 16 clips correctly identified and three Pontianak Malay clips misclassified as Riau Malay. For Pontianak Malay, precision, recall, and F1-score were 0.8667, 0.8125, and 0.8387, respectively, with 13 of 16 clips correctly identified and three clips misclassified as Riau Malay. Overall, the macro-average precision, recall, and F1-score were 0.8451, 0.8438, and 0.8436. Because both classes contained 16 test clips, the macro and weighted averages were identical. These results indicate relatively balanced performance within the current grouped held-out test set but do not confirm external generalisation. The detailed classification outcomes are presented in the confusion matrix in Figure 6.

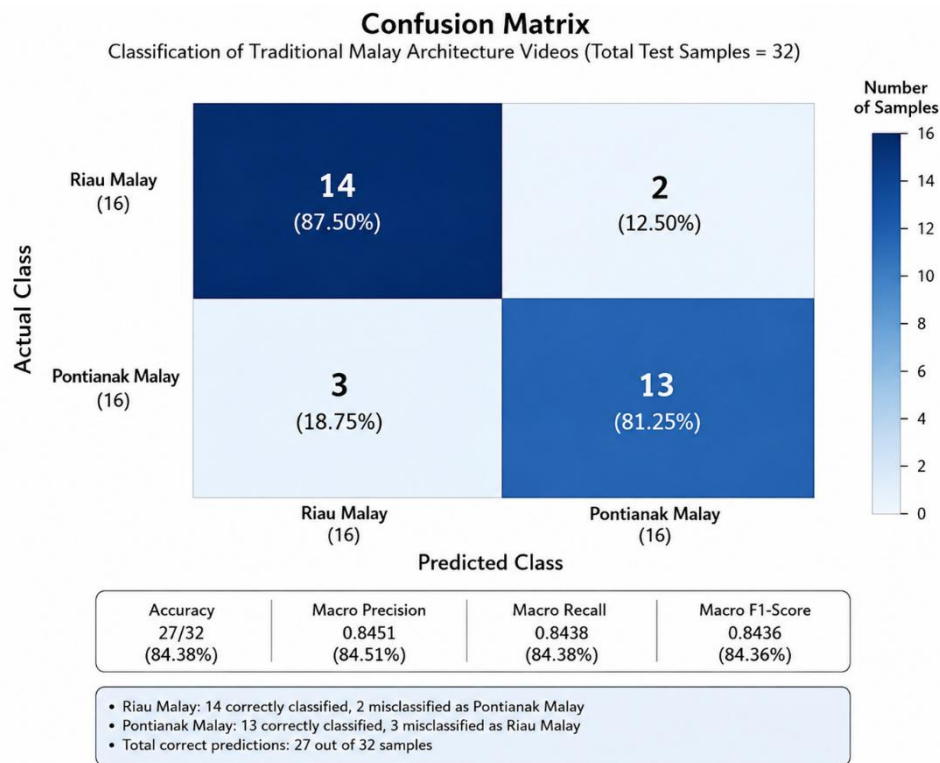


Figure 6. Confusion Matrix of the Proposed Vision Transformer (ViT) Model for Classifying Riau Malay and Pontianak Malay Traditional Houses

Based on Figure 6, the model correctly classified 14 of 16 Riau Malay samples (87.50%) and 13 of 16 Pontianak Malay samples (81.25%). Overall, 27 of 32 samples were correctly classified, yielding an accuracy of 84.38%, with macro-average precision, recall, and F1-score of 84.51%, 84.38%, and 84.36%, respectively. The higher recall for Riau Malay (0.8750) than Pontianak Malay (0.8125), together with two Riau and three Pontianak misclassifications, suggests that some samples shared similar visual characteristics across the two regional architectural domains.

One possible source of confusion is the presence of visually similar ornamental patterns on roofs, windows, and façades across Malay architectural traditions. Decorative elements may be particularly difficult to distinguish when their differences are represented by small motifs or local variations in texture. Zhou et al. [12] similarly treated decorative-pattern recognition as a fine-grained classification problem rather than as a general building classification problem. Their findings emphasise that decorative patterns contain detailed cultural and historical information that requires models capable of preserving local visual details while also interpreting their broader compositional context. From this perspective, the misclassification results in the present study may not simply represent random prediction errors. They may indicate that the available frames did not consistently expose the most discriminative regional features, particularly when architectural elements were partially visible or recorded from less informative viewpoints. Future analysis should therefore examine class activation or attention maps to determine whether the model focuses on

culturally meaningful regions, such as roof ornamentation, gable decoration, window framing, staircase details, and façade composition.

ViT can support digital documentation by recognising visual features of traditional Malay buildings and generating metadata for inventory, archival search, and typological classification. However, its predictions cannot replace expert assessment because cultural value also depends on historical, functional, material, social, and authenticity considerations. Although ViT-B/16 captures global relationships through self-attention, this study did not compare it with CNN models under identical conditions; therefore, no claim of superiority can be made. Zhou et al. [12] showed that ConvNeXt-L effectively combines local feature extraction with a Transformer-inspired design to recognise fine-grained architectural ornaments. Future studies should compare ViT-B/16, ResNet50, ConvNeXt-L, and CNN–Transformer hybrids using the same data split, augmentation, and evaluation protocol to assess the roles of local texture, global composition, and temporal information.

Nevertheless, the similarity among the three accuracy curves should not be interpreted as conclusive evidence that the model is free from overfitting or that it will generalise equally well to independent architectural data. The evaluation subsets were derived from the same overall collection procedure, geographical domains, and recording conditions. Consequently, the visual distributions of the training, validation, and testing data may remain relatively similar even when the samples are placed in separate subsets.

Although the dataset contained 13,230 extracted frames, these frames were obtained from only 24 source videos. Consecutive frames sampled at five frames per second frequently contained similar architectural content, camera positions, illumination conditions, and backgrounds. Therefore, 13,230 frames should not be interpreted as 13,230 fully independent architectural observations. The effective number of independent observations is more closely related to the number of source videos, buildings, and spatiotemporal segments than to the total number of extracted frames.

The distribution of architectural features was unequal, with full-building and window views more common than carvings and ornaments, which appeared in only two videos. Therefore, feature-specific results may partly reflect differences in data availability and should be interpreted cautiously. The predefined viewpoint and ROI categories improved segmentation consistency but may also have influenced which visual information was retained.

The reported performance is limited to a single grouped hold-out configuration and does not confirm generalisation to external data. Future evaluations should use grouped five-fold cross-validation and include unseen buildings, different recording devices, varied lighting conditions, and cross-regional or cross-dataset testing.

4. CONCLUSION

This study explored the application of a Vision Transformer to recognise visual features of Riau Malay architecture, using Pontianak Malay architecture as a cross-regional comparison. The experiments considered roof shapes, building structures, carvings and ornaments, windows, staircases, and full-building representations. The reported scores varied across architectural features and viewing configurations. Window, staircase, roof, and full-building representations generally produced higher recognition performance when the architectural elements were clearly visible. The classification model employed a ViT-B/16 architecture initialised with ImageNet-1K pretrained weights. However, evaluation on a balanced grouped held-out test set of 32 clips produced an overall accuracy of 84.38%, a macro precision of 84.51%, a macro recall of 84.38%, and a macro F1-score of 84.36%. The model correctly classified 14 of 16 Riau Malay clips and 13 of 16 Pontianak Malay clips. The findings should be interpreted as preliminary because the evaluation was based on a single grouped hold-out configuration and a limited number of source videos. Although grouping was used to prevent frames from the same source segment from being distributed across different subsets, the resulting performance may still depend on the composition of the selected data partition. The present study demonstrates the potential use of Vision Transformer representations for architectural feature recognition and digital heritage documentation. Nevertheless, it does not establish the superiority of ViT over CNN-based architectures or validate the completeness of a digital heritage preservation system. Future work should apply grouped five-fold cross-validation, independent building-level testing, CNN baseline comparisons, ROC–AUC analysis, and attention-based visualisation.

ACKNOWLEDGEMENTS

We want to thank the Ministry of Higher Education, Science, and Technology (Kemdiktisaintek) of the Republic of Indonesia for supporting this research through the 2025 grant funding for the dissertation research scheme, Contract No. 0498.12/LL5-INT/AL.04/2025, 114/PDD/LPPM.UAD/VI/2025. We would also like to thank the Research and Community Service Institute of Ahmad Dahlan University for its guidance from the drafting of the report through to the final report of this research, as well as for the achievement of the mandatory outputs promised in this research.

REFERENCES

- [1] S. S. Workflow, R. Salah, N. Géczy, and K. A. Károlyfi, "Architectural Heritage Digitization : A Classification-Driven," *Buildings*, vol. 16, no. 21, pp. 1–24, 2026, doi: <https://doi.org/10.3390/buildings16010021>.
- [2] F. Colace, R. Gaeta, A. Lorusso, M. Pellegrino, and D. Santaniello, "New AI challenges for cultural heritage protection : A general overview," *J. Cult. Herit.*, vol. 75, pp. 168–193, 2025, doi: [10.1016/j.culher.2025.07.019](https://doi.org/10.1016/j.culher.2025.07.019).
- [3] X. Li and F. Chiabrando, "Machine Learning and Deep Learning for Cultural Heritage Conservation : A Bibliometric and Task-Oriented Review," *Remote Sens.*, vol. 18, no. 268, 2026, doi: <https://doi.org/10.3390/rs18040628>.
- [4] 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](https://doi.org/10.3390/su16114773).
- [5] K. Siountri and C. N. Anagnostopoulos, "The Classification of Cultural Heritage Buildings in Athens Using Deep Learning Techniques," *Heritage*, vol. 6, no. 4, pp. 3673–3705, 2023, doi: [10.3390/heritage6040195](https://doi.org/10.3390/heritage6040195).
- [6] A. Sasithradevi, B. Chanthini, T. Subbulakshmi, and P. Prakash, "MonuNet : a high performance deep learning network for Kolkata heritage image classification," *Herit. Sci.*, pp. 1–14, 2024, doi: [10.1186/s40494-024-01340-z](https://doi.org/10.1186/s40494-024-01340-z).
- [7] A. Luiz and C. Ottoni, "ImageOP : The Image Dataset with Religious Buildings in the World Heritage Town of Ouro Preto for Deep Learning Classification," *Heritage*, vol. 7, no. 11, pp. 6499–6525, 2024, doi: <https://doi.org/10.3390/heritage7110302>.
- [8] Z. Li *et al.*, "A deep learning-based method for identifying traditional villages across cultural types," *Herit. Sci.*, vol. 13, no. 671, pp. 1–15, 2025, doi: <https://doi.org/10.1038/s40494-025-02187-8>.
- [9] T. Luo, X. Sun, W. Zhao, W. Li, L. Yin, and D. Xie, "Ethnic Architectural Heritage Identification Using Low-Altitude UAV Remote Sensing and Improved Deep Learning Algorithms," *Buildings*, vol. 15, no. 15, 2025, doi: <https://doi.org/10.3390/buildings15010015>.
- [10] 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](https://doi.org/10.3390/buildings15111843).
- [11] P. Han, S. Hu, and R. Xu, "Formal Feature Identification of Vernacular Architecture Based on Deep Learning—A Case Study of Jiangsu Province, China," *Sustain.*, vol. 17, no. 4, pp. 1–30, 2025, doi: [10.3390/su17041760](https://doi.org/10.3390/su17041760).
- [12] J. Zhou, L. Xie, P. Fricker, and K. Liu, "ConvNeXt-L-Based Recognition of Decorative Patterns in Historical Architecture : A Case Study of Macau," *Buildings*, vol. 15, no. 20, pp. 1–27, 2025, doi: <https://doi.org/10.3390/buildings15203705>.
- [13] 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, pp. 1–19, 2024, doi: [10.3390/su16114773](https://doi.org/10.3390/su16114773).
- [14] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies*, vol. 13, no. 1, pp. 1–32, 2025, doi: [10.3390/technologies13010032](https://doi.org/10.3390/technologies13010032).
- [15] J. Montrezol, H. S. Oliveira, and H. P. Oliveira, "Decoding Vision Transformer Variations for Image Classification: A Guide to Performance and Usability," *Machine Learning with Applications*, vol. 23, p. 100844, Mar. 2026, doi: [10.1016/j.mlwa.2026.100844](https://doi.org/10.1016/j.mlwa.2026.100844).
- [16] W. Xu *et al.*, "ViT-HVE : a vision transformer-based framework for recognition and weighted evaluation of cultural heritage values," *Herit. Sci.*, vol. 13, no. 571, 2025, doi: <https://doi.org/10.1038/s40494-025-02143-6>.
- [17] T. Fan and H. Wang, "Multimodal cultural heritage image recognition based on quantum and classical multimodal fusion network," *Herit. Sci.*, vol. 14, p. 160, 2026, doi: <https://doi.org/10.1038/s40494-026-02419-5>.
- [18] S. Khan, M. Naseer, M. Hayat, and S. W. Zamir, "Transformers in Vision : A Survey," *ACM Comput. Surv.*, pp. 1–30, 2021, doi: <https://doi.org/10.1145/3505244>.
- [19] T. Luo, X. Sun, W. Zhao, W. Li, L. Yin, and D. Xie, "Ethnic Architectural Heritage Identification Using Low-Altitude UAV Remote Sensing and Improved Deep Learning Algorithms," *Buildings*, vol. 15, no. 1, 2025, doi: [10.3390/buildings15010015](https://doi.org/10.3390/buildings15010015).
- [20] 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.*, no. May, 2024, doi: [10.1109/ICIESTR60916.2024.10798341](https://doi.org/10.1109/ICIESTR60916.2024.10798341).

[21] A. Gil, Y. Arayici, B. Kumar, and R. Laing, "Machine and Deep Learning Implementations for Heritage Building Information Modelling : A Critical Review of Theoretical and Applied," *ACM J. Comput. Cult. Herit.*, vol. 17, no. 3, 2024, doi: <https://doi.org/10.1145/3649442>.

[22] A. Reshetnikov *et al.*, "DEArt : Building and evaluating a dataset for object detection and pose classification for European art," *J. Cult. Herit.*, vol. 75, pp. 258–266, 2025, doi: 10.1016/j.culher.2025.07.022.

[23] A. M. Tasir, M. Nor, and A. Khalid, "OPEN A novel deep neural model for efficient and scalable historical place image classification," *Sci. Rep.*, vol. 15, no. 42745, pp. 1–29, 2025, doi: <https://doi.org/10.1038/s41598-025-26897-y>.

[24] A. M. Shehata and N. S. Alaboud, "Identification of Features of Architectural Heritage Using Deep Learning Techniques," *Civ. Eng. Archit.*, vol. 13, no. 4, pp. 3280–3298, 2025, doi: 10.13189/cea.2025.130431.

[25] D. W. Alexey Dosovitskiy, Lucas Beyer*, Alexander Kolesnikov*, M. M. Xiaohua Zhai*, Thomas Unterthiner, Mostafa Dehghani, and N. H. Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Published as a conference paper at ICLR 2021*, 2021.

[26] A. Dosovitskiy *et al.*, "An image is worth 16 x 16 words :," *Int. Conf. Learn. Represent.*, pp. 1–21, 2021, doi: <https://doi.org/10.48550/arXiv.2010.11929>.

[27] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 6816–6826, 2021, doi: 10.1109/ICCV48922.2021.00676.

[28] T. Zhang, W. Xu, B. Luo, and G. Wang, "Depth-Wise Convolutions in Vision Transformers for Efficient Training on Small Datasets," *Neurocomputing*, vol. 617, p. 128998, Nov. 2024, doi: 10.1016/j.neucom.2024.128998.

[29] E. Matsuyama, H. Watanabe, and N. Takahashi, "Performance Comparison of Vision Transformer- and CNN-Based Image Classification Using Cross Entropy : A Preliminary Application to Lung Cancer Discrimination from CT Images," *J. Biomed. Sci. Eng.*, vol. 17, no. 9, pp. 157–170, 2024, doi: <https://doi.org/10.4236/jbise.2024.179012>.

BIOGRAPHIES OF AUTHORS

Heri Pramono	Heri Pramono is a lecturer at STAR YKPN Yogyakarta and is active in digital architecture. He has experience in teaching, AI-based research, and BIM to support digital architecture. His academic focus includes architectural design studios, building physics, Building Information Modeling, and digital architecture.
Sri Winiarti	Sri Winiarti is a lecturer in Informatics at Ahmad Dahlan University (UAD) in Yogyakarta. She specialises in artificial intelligence, data mining, and intelligent systems. Active in research and scientific publications, she has largely focused on applying AI to problem-solving across various domains. Her primary research areas include Artificial Intelligence, Research Methodology, Databases, and Decision Support Systems. Her scientific publications, which align with her expertise in AI, have been published both nationally and internationally.
Abdul Fadlil	Abdul Fadlil is a lecturer in the Doctoral Program in Informatics at UAD Yogyakarta. He is known for his active involvement in artificial intelligence, image processing, and intelligent systems. In addition to teaching, he is also involved in various strategic research projects and has published scientific papers at the national and international levels.
Sunardi	Sunardi is a lecturer in the Doctoral Program in Informatics at UAD Yogyakarta, specialising in intelligent systems and information technology. He is actively involved in research, technology development, and scientific publications, particularly in applying AI and data analysis to a range of industrial and academic needs.