frontiers | Frontiers in Neurorobotics

TYPE Original Research
PUBLISHED 13 December 2024
DOI 10.3389/fnbot.2024.1513488

OPEN ACCESS

EDITED BY
Xin Jin,
Yunnan University, China

REVIEWED BY
Jie Zhang,
Xi'an University of Technology, China
Peijie Qiu,
Washington University in St. Louis,
United States

*CORRESPONDENCE
Yu Caixia
✉ yodb32@163.com

RECEIVED 18 October 2024
ACCEPTED 18 November 2024
PUBLISHED 13 December 2024

CITATION
Yongyin L and Caixia Y (2024) Cross-attention swin-transformer for detailed segmentation of ancient architectural color patterns.
Front. Neurorobot. 18:1513488.
doi: 10.3389/fnbot.2024.1513488

Cross-attention swin-transformer for detailed segmentation of ancient architectural color patterns

Lv Yongyin and Yu Caixia*

Department of Fine Arts, Bozhou University, Bozhou, Anhui, China

Introduction: Segmentation tasks in computer vision play a crucial role in various applications, ranging from object detection to medical imaging and cultural heritage preservation. Traditional approaches, including convolutional neural networks (CNNs) and standard transformer-based models, have achieved significant success; however, they often face challenges in capturing fine-grained details and maintaining efficiency across diverse datasets. These methods struggle with balancing precision and computational efficiency, especially when dealing with complex patterns and high-resolution images.

Methods: To address these limitations, we propose a novel segmentation model that integrates a hierarchical vision transformer backbone with multi-scale self-attention, cascaded attention decoding, and diffusion-based robustness enhancement. Our approach aims to capture both local details and global contexts effectively while maintaining lower computational overhead.

Results and discussion: Experiments conducted on four diverse datasets, including Ancient Architecture, MS COCO, Cityscapes, and ScanNet, demonstrate that our model outperforms state-of-the-art methods in accuracy, recall, and computational efficiency. The results highlight the model's ability to generalize well across different tasks and provide robust segmentation, even in challenging scenarios. Our work paves the way for more efficient and precise segmentation techniques, making it valuable for applications where both detail and speed are critical.

KEYWORDS

segmentation, vision transformer, multi-scale attention, robustness enhancement, computational efficiency

1 Introduction

The task of segmenting color patterns in ancient architectural images is crucial for the preservation, restoration, and study of historical artifacts (Minaee et al., 2021). These intricate designs not only reflect the cultural and artistic achievements of past civilizations but also carry significant historical and symbolic meanings. Accurate segmentation allows for detailed analysis and digital archiving, which helps in restoration efforts and ensures the longevity of cultural heritage (Lin et al., 2024b). Additionally, such segmentation facilitates virtual reconstruction and tourism, enabling broader accessibility to these artifacts. However, this task is challenging due to the complexity of the patterns, variations in color and texture, and potential degradation over time (Yanowitz and Bruckstein, 1989). Therefore, there is a need for advanced computational techniques to accurately segment and analyze these patterns, supporting heritage conservation and academic research.

To address the limitations of manual segmentation and traditional pattern recognition techniques, early research in this field focused on symbolic AI and knowledge-based approaches. These methods relied on predefined rules and symbolic representations to identify specific motifs and color distributions present in ancient architectural designs (Abdullah et al., 2023). The core idea was to encode expert knowledge, often derived from historians, architects, or cultural heritage experts, into a set of rules that could be programmed into the system. For instance, these rules could specify the geometric shapes, color palettes, or symmetry patterns that are commonly observed in traditional architectures, allowing the system to segment and classify these elements accordingly. This knowledge-based approach could effectively handle simple and repetitive patterns, such as borders, uniform motifs, or repetitive floral designs, making them particularly suitable for controlled environments where variations were minimal (Lin et al., 2024a). Moreover, symbolic AI systems were relatively transparent and interpretable, allowing experts to understand how decisions were being made by the algorithm. This made it easier to debug or refine the system, as the rules could be adjusted or expanded based on new insights (Bennetot et al., 2022). However, these systems were often rigid and struggled to adapt to variations in pattern design or color caused by environmental factors such as lighting, aging, and deterioration of the material. For example, an image captured under different lighting conditions might cause colors to appear different, making it difficult for rule-based systems to maintain accuracy. Additionally, aging and wear can lead to faded colors or partially obscured patterns, which would not fit neatly into the predefined rules, resulting in misclassifications or incomplete segmentations. Another major drawback of these knowledge-based systems was their scalability. Developing and refining rules required extensive domain expertise and manual effort, which limited their ability to be applied across diverse datasets. Each new dataset or variation of pattern would potentially require a re-evaluation of the rules or the addition of new ones, leading to increased costs and time (Hong et al., 2024). While these methods were foundational in introducing automated segmentation, their inability to generalize across different datasets or adapt to new scenarios effectively made them less viable for more extensive, real-world applications. This rigidity and dependence on predefined rules prompted the exploration of more flexible, data-driven approaches that could learn directly from the data itself, without the need for explicit rule-coding, paving the way for machine learning techniques.

To overcome the rigidity of rule-based systems, researchers began to explore data-driven and machine learning techniques that could learn patterns directly from the data without the need for explicit programming of rules. This shift marked a significant evolution in segmentation tasks, as it allowed for more flexible models that could adapt to variations within the data. Machine learning models, which include classical algorithms such as Support Vector Machines (SVMs) (Ai et al., 2023), k-Nearest Neighbors (k-NN) (Fuadah et al., 2020), and Random Forests (Dhivyaa et al., 2020), rely on statistical learning to identify patterns within a set of labeled data. These models operate by training on features extracted from the images, such as edges, textures, and color histograms, to distinguish between different classes or elements of the architectural patterns. The advantage of these methods over symbolic AI was their ability to adapt to new data by learning from examples, which meant they could handle more variation in the input images. For instance, an SVM could be trained to recognize different color distributions across various lighting conditions, or a Random Forest could classify textures even if the patterns were slightly worn or degraded (Wu et al., 2022). This adaptability made machine learning approaches more robust in dealing with real-world data, which is often subject to inconsistencies and noise. However, the performance of these models heavily depended on the quality of the extracted features, which still required domain knowledge and manual design (Kheradmandi and Mehranfar, 2022). Feature extraction processes such as edge detection or texture mapping needed to be carefully crafted to ensure that the most relevant information was captured, which introduced a degree of subjectivity and potential bias.

With the advent of deep learning and pre-trained models, a new era of segmentation techniques emerged, significantly improving the accuracy and robustness of pattern recognition in ancient architectural images. Deep learning models, particularly Convolutional Neural Networks (CNNs) (Sultana et al., 2020), can learn features directly from the data, eliminating the need for manual feature engineering. These models have been applied to various tasks, including semantic segmentation, where they have demonstrated superior performance due to their ability to learn hierarchical representations of data. Further advancements with architectures like Fully Convolutional Networks (FCNs) (Calisto and Lai-Yuen, 2020), U-Net (Siddique et al., 2021), and DeepLab (Azad et al., 2022) allowed for pixel-level segmentation, enabling precise delineation of patterns. More recently, transformer-based models and pre-trained architectures have been employed to handle long-range dependencies and complex spatial relationships, overcoming the limitations of earlier machine learning models. Despite their success, deep learning approaches require large datasets for training, and their performance can be hindered by variations in data quality, such as noise and degradation present in historical artifacts. Furthermore, these models are often computationally intensive, making them less feasible for real-time applications or deployment on devices with limited resources.

To address the above limitations, we propose our method ArchPaint-Swin, which combines the strengths of hierarchical vision transformers and multi-scale attention mechanisms to effectively segment complex color patterns in ancient architectural images. Unlike traditional models that either rely on rigid rules or manual feature design, ArchPaint-Swin leverages the power of deep learning while integrating multi-scale processing and adaptive feature refinement, ensuring robust performance even under challenging conditions. By incorporating diffusion-based robustness enhancement, our model can handle variations in color, texture, and quality, making it well-suited for the segmentation of intricate historical patterns. This approach not only improves accuracy and efficiency but also reduces the computational overhead, making it viable for both large-scale analysis and real-time applications in the field of cultural heritage preservation.

- Our method introduces a new hierarchical vision transformer backbone combined with multi-scale self-attention, enabling precise segmentation of intricate patterns by effectively capturing both local and global features.

- The approach is adaptable across multiple scenarios, ensuring high efficiency and generalizability, making it suitable for diverse datasets and real-time applications with reduced computational requirements.
- Experimental results demonstrate that our model consistently outperforms state-of-the-art methods, achieving higher accuracy, recall, and F1 scores while maintaining lower inference time and computational overhead.

2 Related work

2.1 Traditional convolutional neural networks for segmentation

Convolutional Neural Networks (CNNs) have been the backbone of many image segmentation models due to their ability to capture spatial hierarchies in images. Early approaches like Fully Convolutional Networks (FCNs) laid the foundation for using deep learning in pixel-wise classification tasks. Subsequent improvements, such as U-Net and its variants, introduced skip connections to merge features from different layers, allowing for the capture of both high-level semantics and low-level details. Models like DeepLab series utilized dilated convolutions to expand the receptive field without increasing computational cost, significantly improving performance on complex datasets like Pascal VOC and MS COCO (Wang, 2024). Despite their success, CNN-based models have limitations, particularly when it comes to handling long-range dependencies within images. This is because convolution operations are inherently local, capturing information from a limited receptive field. To mitigate this, techniques such as atrous convolutions and multi-scale feature pyramids have been used. However, these solutions often come at the cost of increased computational overhead and fail to capture relationships effectively across distant regions (Jin et al., 2024b). This limitation has prompted the exploration of alternative architectures, such as transformers, which can handle global contexts more naturally. Our work seeks to address the limitations of traditional CNNs by leveraging hierarchical vision transformers that combine multi-scale features with efficient attention mechanisms (Li et al., 2024).

2.2 Vision transformers and their applications in segmentation

2.2.1 Transformer-based UNet variants

Swin-UNet (Cao et al., 2022) is a pioneering model that combines Swin Transformer with the UNet structure, achieving remarkable results in medical image segmentation. It employs a sliding window approach to effectively capture both local and global features. However, the basic structure of Swin-UNet has since been extended by numerous variants that further enhance its multi-scale processing and detail preservation capabilities. For instance, Rahman et al. introduced the Multi-scale Hierarchical Vision Transformer (Rahman and Marculescu, 2024), which utilizes cascaded attention decoding to achieve superior fine-grained segmentation. Similarly, Xie Y. et al. (2021) proposed CoTr, which efficiently bridges CNN and Transformer layers to improve segmentation performance in 3D medical imaging. Furthermore, AgileFormer (Qiu et al., 2024) incorporates spatially adaptive Transformer modules, enabling the model to adjust dynamically to input images of different resolutions. In comparison, our model advances multi-scale feature extraction and information flow with unique adaptive multi-scale attention and cascaded attention decoding mechanisms, achieving robustness and precision in segmenting intricate architectural and cultural heritage images.

2.2.2 Skip-connection-based models

Many segmentation models leverage skip-connections to enhance detail preservation and cross-scale feature fusion. UC-TransUNet (Wang et al., 2022), for example, redefines the skip-connections of UNet from a channel-wise perspective by introducing Transformer modules, thereby capturing multi-scale contextual information more effectively. UNet 3+ (Huang et al., 2020) further improves feature propagation across scales through full-scale connections, enabling low- and high-resolution features to be fused more efficiently within the network. Attention U-Net (Oktay, 2018) incorporates attention mechanisms that automatically focus on more informative regions of the image, improving segmentation accuracy. While these models improve in terms of detail handling and boundary preservation, our model's cascaded attention decoding incorporates a dynamic feedback mechanism, allowing it to adaptively process details across scales. This capability is particularly advantageous for segmenting architectural images with complex details, a feature that has not been extensively explored in existing models.

2.2.3 Diffusion models for segmentation

Diffusion models have recently shown promise in image generation and segmentation tasks. SegDiff (Amit et al., 2021), for instance, proposes a segmentation approach based on diffusion probabilistic models, progressively removing noise to obtain refined segmentation results. Although diffusion models demonstrate robust performance in segmentation tasks, their applications are primarily focused on generative tasks. In our work, we innovatively incorporate a diffusion model within the segmentation pipeline as a conditional adaptive diffusion module. This module dynamically adjusts the diffusion strength based on contextual information from the image, a feature that distinguishes it from conventional diffusion-based post-processing (Ruiying, 2024). This conditional adaptive diffusion process provides our model with enhanced robustness, allowing it to handle noise, degradation, and other artifacts commonly encountered in architectural and cultural heritage images. By integrating the diffusion model directly within the segmentation process, we enable more effective noise management, which is critical in preserving details and improving segmentation accuracy in challenging real-world scenarios (Jin et al., 2024a).

2.3 Robustness enhancement and multi-scale feature fusion in segmentation

Robust segmentation models need to handle diverse and noisy data, especially in real-world applications where conditions can vary significantly. Traditional approaches often rely on data augmentation and ensemble methods to improve robustness. However, these methods increase the complexity of the training process and do not guarantee improved performance in unseen scenarios. Recent research has focused on incorporating robustness directly into the model architecture. For instance, the use of attention mechanisms helps models focus on relevant features, but these mechanisms are still sensitive to noise and may struggle to generalize across different datasets (Atzori et al., 2016). Multi-scale feature fusion has been a popular technique to address these issues. Methods like Feature Pyramid Networks (FPN) combine features extracted at different scales to capture both fine details and broader contextual information. However, simple feature merging is often insufficient, as it can lead to redundant information and inefficient processing. More recent works, such as HRNet, attempt to maintain high-resolution representations throughout the network, enabling better feature preservation (Jin et al., 2023). Our approach extends this idea by introducing a dynamic feature fusion strategy that selectively integrates features across scales, guided by adaptive attention mechanisms. Additionally, we enhance robustness by incorporating a diffusion-based module, which iteratively refines features to reduce noise and improve the quality of segmentation outputs. This design allows our model to achieve superior performance across diverse datasets, making it more resilient in challenging environments and more efficient in real-time applications.

3 Methodology

3.1 Overview

Our proposed model introduces a novel approach to segmentation tasks by integrating multiple advanced architectures and strategies to enhance performance in complex scenarios. The model primarily leverages the synergy between multi-scale hierarchical structures, robust fusion mechanisms, and attention-driven modules to achieve high-resolution, accurate segmentation results across diverse data types. The core structure comprises a hierarchical multi-stage processing pipeline, designed to capture fine-grained details as well as broad contextual information, ensuring that both local and global features are effectively utilized. The architecture initiates with a multi-scale transformer-based backbone that performs hierarchical processing, capturing features across various resolutions and scales. By embedding multiple attention windows, it effectively addresses the limitations of traditional single-scale attention models. This is followed by a sophisticated fusion mechanism that dynamically integrates features from different scales, ensuring consistency and reducing redundancy. Subsequently, an advanced attention-based decoder refines these multi-modal features to produce highly accurate segmentation maps, while a diffusion-based enhancement module ensures robustness against variations and improves the generalization capability.

Our model integrates several innovative components to handle the challenges inherent in ancient architectural images, particularly those affected by degradation, noise, and artifacts. Central to this is the diffusion-based robustness enhancement module, which iteratively refines feature representations to reduce noise while preserving intricate details. This module adapts its denoising strength contextually: it applies gentler processing to areas with delicate patterns, ensuring key features are maintained, and more intensive denoising in smoother regions to clear away irrelevant noise. This adaptive approach helps the model retain crucial architectural details while minimizing the impact of fading and other common forms of image degradation. The model also employs a multi-scale attention mechanism to capture both local and global features, enhancing resilience to lighting inconsistencies and quality variations. By analyzing information across multiple scales, the model isolates essential structural details from common artifacts, such as scratches and discolorations. Additionally, a dynamic feature feedback loop within the diffusion module refines feature representations iteratively, allowing the model to focus on significant architectural elements while progressively filtering out noise. Together with data augmentation techniques that simulate real-world variations, such as blur and lighting shifts, and positional encoding to maintain spatial consistency, these strategies enable the model to effectively analyze and segment ancient architectural patterns even in challenging conditions. This comprehensive approach makes the model robust and highly suited for the digital preservation of cultural heritage images.

We organize this section as follows. In Section 3.2, we describe the hierarchical vision transformer backbone that is at the core of our model, explaining how multi-scale attention is achieved. Section 3.3 discusses the fusion mechanism and the innovative cascaded attention decoding strategy that aggregates features efficiently. Finally, in Section 3.4, we introduce the diffusion-based enhancement module, which mitigates noise and improves the robustness of the segmentation output. Through this modular and highly integrated design, our model aims to set a new benchmark in segmentation tasks, effectively combining state-of-the-art techniques to address challenges in precision, robustness, and computational efficiency.

3.2 Hierarchical vision transformer backbone

The backbone of our model is constructed upon a hierarchical vision transformer framework, designed to capture multi-scale features across different levels of resolution. Unlike traditional transformer architectures that rely on a single-scale self-attention mechanism, our approach implements a multi-scale self-attention strategy, allowing the model to process visual information at various granularities. This method addresses the inherent limitations of single-scale attention by enabling the model to learn more generalizable features, thus improving performance across diverse segmentation tasks (as shown in Figure 1).

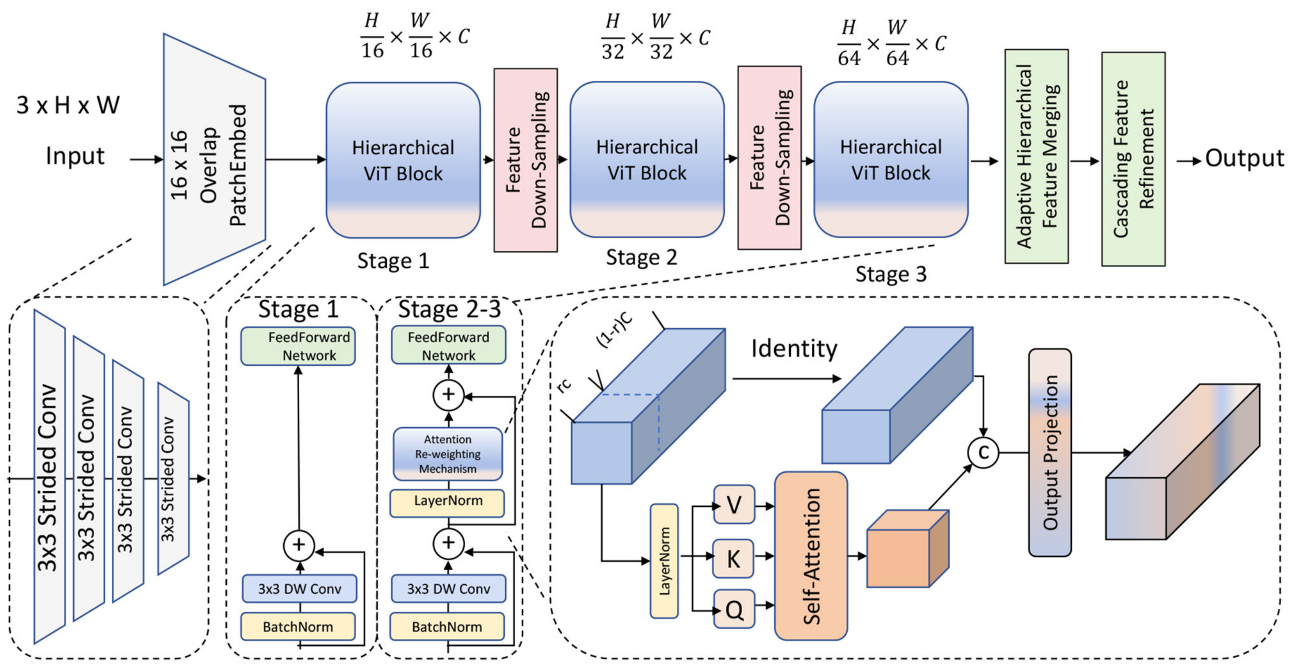

FIGURE 1
A hierarchical vision transformer model architecture. The model begins with an overlapping patch embedding module, followed by three main stages, each containing multi-scale self-attention blocks with feature down-sampling to capture different granularities. The architecture includes adaptive hierarchical feature merging and cascading feature refinement to integrate information across scales, enhancing fine-grained details and broader contextual understanding. The output layer consolidates these features, optimized by an attention re-weighting mechanism for improved segmentation accuracy.

3.2.1 Multi-scale self-attention module

The core of the hierarchical backbone in our model is the multi-scale self-attention module, denoted as $\mathcal{A}_{ms}$. This module is designed to address the limitations of conventional attention mechanisms by enabling the model to capture features across multiple scales, effectively combining fine-grained local details with broader contextual information. Let $\mathbf{I}$ be the input image, which is initially divided into smaller patches. Each patch is processed through a stem network, which consists of a series of convolutional layers that transform the raw pixel data into a set of feature embeddings $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n\}$. These embeddings represent the essential characteristics of the patches and serve as inputs to the multi-level transformer blocks.

The architecture is designed to handle multiple hierarchical levels, where each level processes the embeddings at different scales. The output of the transformer block at a specific level l is denoted by $\mathbf{F}_l$, which can be mathematically expressed as:

$$\mathbf{F}_l = \mathcal{A}_{ms}(\mathbf{E}_l) = \text{Softmax}\left(\frac{\mathbf{Q}_l\mathbf{K}_l^\top}{\sqrt{d_k}}\right)\mathbf{V}_l, \qquad (1)$$

where $\mathbf{E}_l$ represents the input embeddings at level l, and $\mathbf{Q}_l, \mathbf{K}_l, \mathbf{V}_l$ are the query, key, and value matrices, respectively, derived from $\mathbf{E}_l$. The matrices are calculated as follows:

$$\mathbf{Q}_l = \mathbf{W}_Q\mathbf{E}_l, \quad \mathbf{K}_l = \mathbf{W}_K\mathbf{E}_l, \quad \mathbf{V}_l = \mathbf{W}_V\mathbf{E}_l, \qquad (2)$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable weight matrices. The attention mechanism computes a weighted combination of values $\mathbf{V}_l$, where the weights are determined by the similarity between queries $\mathbf{Q}_l$ and keys $\mathbf{K}_l$, scaled by the dimensionality d_k to stabilize gradients during training. This scaling can be particularly important when working with large-scale data, as it prevents the computed values from becoming excessively large.

The multi-scale aspect is crucial in ensuring that both fine and coarse features are captured simultaneously. To achieve this, the feature embeddings are processed at multiple resolutions within the self-attention module. Specifically, we modify the input embeddings by applying a series of down-sampling and up-sampling operations, allowing the model to adjust the scale at which the attention is computed. For example, at a higher level, the embeddings might be down-sampled, focusing on broader patterns and contextual information, whereas at a lower level, the embeddings might maintain a higher resolution, preserving fine-grained details. The multi-resolution processing can be described by:

$$\mathbf{F}_l^{(s)} = \text{Softmax}\left(\frac{\mathbf{Q}_l^{(s)}\mathbf{K}_l^{(s)\top}}{\sqrt{d_k}}\right)\mathbf{V}_l^{(s)}, \qquad (3)$$

where s denotes different scales, and $\mathbf{Q}_l^{(s)}, \mathbf{K}_l^{(s)}, \mathbf{V}_l^{(s)}$ are the scaled versions of the query, key, and value matrices. The outputs from various scales are then aggregated to form a comprehensive representation:

$$\mathbf{F}_l = \sum_{s=1}^{S} \alpha_s \mathbf{F}_l^{(s)}, \qquad (4)$$

where α_s are learnable weights that determine the contribution of each scale. This multi-scale aggregation allows the model to effectively integrate information from different resolutions, enhancing its ability to discern intricate patterns that might otherwise be missed. Additionally, to further improve the robustness of feature extraction, a positional encoding is added to the input embeddings $\mathbf{E}_l$ to preserve spatial information, ensuring that the attention mechanism accounts for the relative positions of patches within the image. The enhanced design of the multi-scale self-attention module enables the model to seamlessly adapt to complex patterns, making it well-suited for tasks requiring detailed segmentation, such as those involving ancient architectural color patterns.

3.2.2 Cascading feature refinement mechanism

To enhance feature aggregation across different scales, we introduce a cascading feature refinement mechanism within the transformer blocks. This mechanism is designed to iteratively improve feature representations by aggregating information from multiple layers, effectively combining low-level details with high-level contextual features. Unlike traditional hierarchical approaches that process features in a one-way manner, the cascading feature refinement allows each block to incorporate information from preceding layers. This enables the model to refine high-level representations without losing the essential context provided by initial, low-level details, leading to more accurate segmentation outputs (Figure 2). Formally, this cascading operation can be expressed as:

$$\mathbf{H}_l = \mathbf{F}_l + \sum_{i=1}^{l-1} \alpha_i \mathbf{F}_i, \tag{5}$$

where $\mathbf{F}_l$ denotes the feature output at the current level l, and $\mathbf{F}_i$ represents the feature maps from previous levels. The coefficients α_i are learnable weights that adaptively adjust the contribution of each preceding feature map, ensuring that only the most relevant information is passed forward. This adaptive weighting makes the refinement process more flexible, as it allows the model to prioritize different feature aspects depending on the data characteristics. The cascading mechanism is particularly beneficial in scenarios where segmentation tasks involve intricate patterns and subtle details, such as in architectural designs. By enabling a multi-level refinement, the model ensures that details are preserved while higher-level features capture broader, more abstract concepts. This results in a more cohesive representation that enhances the overall segmentation accuracy.

3.2.3 Adaptive hierarchical feature merging strategy

A critical component of the backbone is the adaptive hierarchical feature merging strategy, which consolidates information from different attention heads and scales. The goal of this strategy is to create a unified feature representation that seamlessly integrates multi-scale features extracted across various layers. This is achieved by using the multi-scale attention modules, $\mathcal{A}_{ms}$, at each hierarchical level, and combining their outputs. Let $\mathbf{F}_{ms}$ be the final combined multi-scale feature map, then it can be defined as:

$$\mathbf{F}_{ms} = \text{Concat}(\mathcal{A}_{ms}^1(\mathbf{E}_1), \mathcal{A}_{ms}^2(\mathbf{E}_2), ..., \mathcal{A}_{ms}^L(\mathbf{E}_L)), \tag{6}$$

where $\mathcal{A}_{ms}^j$ represents the attention mechanism applied at the j-th level, and L denotes the total number of hierarchical levels. Each $\mathcal{A}_{ms}^j$ processes features at different resolutions, allowing the model to extract both coarse and fine details. The concatenation operation combines these multi-scale features, which are then passed through a set of convolutional layers to ensure smooth integration.

To further enhance the spatial consistency of the output, the concatenated feature map undergoes a series of convolutional and normalization layers that align and smooth out discrepancies between features from different scales. Let the processed output be represented as:

$$\mathbf{F}_{final} = \text{Conv}(\mathbf{F}_{ms}) + \gamma \cdot \text{Norm}(\mathbf{F}_{ms}), \tag{7}$$

where $\text{Conv}(\cdot)$ denotes the convolutional operations, $\text{Norm}(\cdot)$ represents batch normalization or layer normalization, and γ is a learnable parameter that adjusts the influence of the normalization. This final adjustment helps in maintaining consistency across spatial dimensions, ensuring that the features can be effectively used in subsequent processing steps. The adaptive hierarchical feature merging strategy is essential for tasks where understanding multi-scale features is critical. By consolidating information across multiple levels, the model ensures that both local patterns and global structures are captured, providing a robust feature set for downstream segmentation tasks. This design not only improves segmentation performance but also makes the model more resilient to variations in input data, such as changes in lighting, scale, or texture. Together, the cascading feature refinement and adaptive hierarchical merging enable the model to deliver high-precision, efficient segmentation results across diverse and challenging datasets.

Moreover, to handle variations in input data and improve generalization, our architecture includes a robust attention re-weighting mechanism, defined as:

$$\mathbf{W}_{att} = \text{Sigmoid}(\mathbf{W}_{init} + \gamma \cdot \Delta\mathbf{W}), \tag{8}$$

where $\mathbf{W}_{init}$ is the initial attention weight, γ is a scaling factor, and $\Delta\mathbf{W}$ represents the deviation observed during training. This adaptive re-weighting allows the model to focus more effectively on relevant regions, even when the input data varies significantly across instances.

3.3 Fusion and cascaded attention decoding

In order to effectively integrate multi-scale features and enhance segmentation accuracy, our model employs a novel Fusion and Cascaded Attention Decoding mechanism. This approach ensures that features from different stages of the backbone are aggregated and refined through a sophisticated decoding process, leading to improved performance, especially in scenarios involving complex visual patterns.

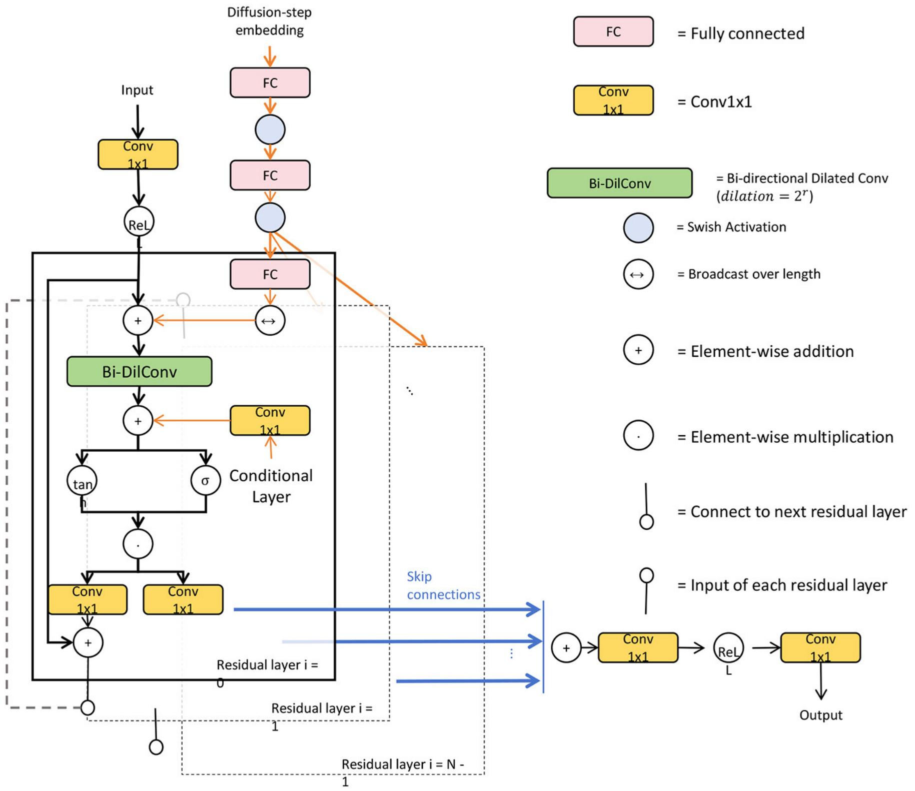

FIGURE 2
A detailed schematic diagram of a bidirectional dilated convolutional network with cascading feature refinement, showcasing the input pathway, diffusion-step embedding, and multiple residual layers. Key elements include Conv1x1 layers, fully connected (FC) layers, bidirectional dilated convolutions, and element-wise operations. The structure highlights skip connections and condition inputs, integrating both temporal and spatial features, while arrows indicate data flow across the network. The model employs Swish activation functions to enhance non-linearity and refine feature representations at each stage for improved accuracy in segmentation tasks.

3.3.1 Dynamic multi-scale feature fusion

The first step in the decoding process is the dynamic multi-scale feature fusion, which integrates feature maps obtained from the hierarchical vision transformer backbone. The backbone produces feature maps at different scales, denoted as $\mathbf{F}_{ms}^{(1)}, \mathbf{F}_{ms}^{(2)}, ..., \mathbf{F}_{ms}^{(L)}$, where L represents the total number of scales. Each feature map captures information at a specific resolution, allowing the model to understand both fine details and broader structures. The fusion mechanism adaptively combines these multi-scale features to produce a unified representation that retains essential information from each scale, leading to more accurate and robust segmentation outcomes.

The fusion process can be mathematically expressed as:

$$\mathbf{F}_{fused} = \sum_{j=1}^{L} \beta_j \cdot \mathrm{Conv}_{1\times1}(\mathbf{F}_{ms}^{(j)}), \tag{9}$$

where β_j are learnable weights that dynamically adjust the contribution of each scale, ensuring that the most relevant information is emphasized. The term $\mathrm{Conv}_{1\times1}$ refers to a 1×1 convolution, which plays a crucial role in aligning the feature dimensions across different scales. This alignment is necessary because feature maps at various scales might have different resolutions, and the 1×1 convolution standardizes these dimensions, allowing them to be effectively integrated.

One of the key advantages of this dynamic fusion process is its ability to adaptively learn which scales are most important for the task at hand. For instance, in scenarios where fine-grained details are critical (e.g., segmenting intricate patterns in architectural images), the model can assign higher weights β_j to lower-scale features that retain these details. Conversely, in tasks that require understanding of broader structures (e.g., identifying large regions in cityscapes), higher-scale features can be prioritized. This adaptability makes the model more versatile and capable of handling a variety of segmentation challenges. Moreover, the fusion process ensures that information from various scales is smoothly combined, reducing inconsistencies that might arise if the scales were processed independently. By summing the outputs from different scales with learnable weights, the model effectively preserves critical details across the image while also maintaining contextual information. The smooth integration is facilitated by the 1×1 convolutions, which not only align the dimensions but also help reduce noise and enhance the most salient features, resulting in a cleaner and more accurate segmentation map.

The effectiveness of the dynamic multi-scale feature fusion can be further enhanced by additional layers of processing after the initial fusion. For example, the fused feature map $\mathbf{F}_{fused}$ can be passed through additional convolutional and normalization layers to further refine the integrated features:

$$\mathbf{F}_{refined} = \text{Norm}(\text{Conv}_{3\times3}(\mathbf{F}_{fused})) + \mathbf{F}_{fused}, \tag{10}$$

where $\text{Conv}_{3\times3}$ refers to a standard convolution operation that captures local relationships, and Norm indicates a normalization layer that helps stabilize the learning process. The addition of $\mathbf{F}_{fused}$ ensures that the original fused information is preserved, similar to a residual connection, which enhances the robustness and stability of the fusion. The dynamic multi-scale feature fusion mechanism ensures that the decoding process can effectively utilize information across various scales, balancing the need for both detailed and contextual understanding. This capability is crucial for tasks like segmenting architectural patterns or urban scenes, where both small details and larger structures need to be accurately identified. By learning to adaptively weigh features from different scales, the model becomes more efficient and accurate, capable of delivering high-quality segmentation results across diverse scenarios.

3.3.2 Cascaded attention decoding

After the multi-scale features are fused, the model applies a cascaded attention decoding strategy to refine these features progressively, stage by stage. This decoding approach is designed to iteratively enhance the quality of the feature maps, ensuring that each refinement stage builds upon the output of the previous one. By incorporating contextual information at every step, the decoder effectively captures both fine details and broader spatial relationships, leading to more accurate segmentation results (as shown in Figure 3).

Let $\mathbf{D}_{t-1}$ be the output from the previous decoding stage, and $\mathbf{C}_t$ be the contextual feature map that contains relevant information from earlier processing layers or external sources. The refined output $\mathbf{D}_t$ at the current stage t is computed using a cascaded attention mechanism, defined as:

$$\mathbf{D}_t = \mathcal{A}_{cascade}(\mathbf{D}_{t-1}, \mathbf{C}_t) = \text{Softmax}\left(\frac{\mathbf{Q}_t\mathbf{K}_t^{\top}}{\sqrt{d_k}}\right)\mathbf{V}_t, \tag{11}$$

where $\mathbf{Q}_t$, $\mathbf{K}_t$, and $\mathbf{V}_t$ are the query, key, and value matrices. These matrices are derived from the features $\mathbf{D}_{t-1}$ and $\mathbf{C}_t$, ensuring that the decoder can attend to both the refined features and the contextual information. Specifically:

$$\mathbf{Q}_t = \mathbf{W}_Q\mathbf{D}_{t-1}, \quad \mathbf{K}_t = \mathbf{W}_K\mathbf{C}_t, \quad \mathbf{V}_t = \mathbf{W}_V\mathbf{C}_t, \tag{12}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable weight matrices that help transform the feature maps into their respective query, key, and value representations. The attention mechanism $\mathcal{A}_{cascade}$ computes a weighted combination of values $\mathbf{V}_t$, where the weights are determined by the similarity between the queries and keys, scaled by the dimensionality d_k.

The cascading nature of this attention mechanism ensures that the model can iteratively refine the features at each decoding stage. By applying attention repeatedly, the decoder can correct errors, sharpen details, and bring in relevant contextual information from the surrounding regions of the image. This is especially useful in scenarios where segmentation boundaries are ambiguous or where intricate patterns require careful disambiguation, as seen in ancient architectural designs or urban scenes with overlapping structures.

Furthermore, the cascaded attention mechanism introduces a form of feedback loop within the decoder, where the output at each stage is used as an input for the next stage. This iterative refinement can be expressed as:

$$\mathbf{D}_t = \mathbf{D}_{t-1} + \gamma_t \cdot \mathcal{A}_{cascade}(\mathbf{D}_{t-1}, \mathbf{C}_t), \tag{13}$$

where γ_t is a learnable parameter that controls the influence of the attention output on the final feature map. This formulation ensures that the model can adaptively refine the features, emphasizing or de-emphasizing certain details based on the learned weights. The use of residual connections, as shown by the addition of $\mathbf{D}_{t-1}$, helps in stabilizing the learning process, preventing issues like vanishing gradients during backpropagation.

Additionally, the decoding process can be further enhanced by incorporating multi-head attention, where multiple independent attention heads are applied simultaneously. Each head captures a different aspect of the feature map, allowing for a more comprehensive understanding of the spatial relationships within the image:

$$\mathbf{D}_t^{\text{multi}} = \text{Concat}(\mathcal{A}^1_{cascade}(\mathbf{D}_{t-1}, \mathbf{C}_t), ..., \mathcal{A}^H_{cascade}(\mathbf{D}_{t-1}, \mathbf{C}_t)), \tag{14}$$

where H denotes the number of attention heads, and $\mathcal{A}^h_{cascade}$ represents the output from the h-th head. The outputs from all heads are concatenated and then processed through a linear transformation to produce the final refined feature map. Multi-head attention enhances the model's ability to capture multiple aspects of the input simultaneously, making it more robust in handling diverse segmentation challenges. The cascaded attention decoding mechanism ensures that the final output maintains high accuracy, even in challenging regions where traditional methods might struggle. By iteratively refining the features and incorporating contextual cues at every stage, the decoder effectively handles complex patterns, subtle textures, and overlapping elements, resulting in high-quality segmentation maps that are well-suited for both detailed and broad applications.

Increasing the number of decoding stages generally improves performance, particularly in terms of segmentation accuracy and boundary precision. Each additional stage allows the model to iteratively refine the feature maps, incorporating contextual information at multiple levels. This leads to more accurate delineation of complex patterns, which is especially beneficial for architectural images where fine-grained details are critical. However, we observed diminishing returns beyond a certain number of stages. For example, moving from 2 to 4 stages led to noticeable improvements in Dice and IoU scores, but the gains from adding a 5th or 6th stage were minimal. Thus, we determined an optimal range for decoding stages that maximizes performance without excessive computational cost. Each decoding stage adds to the model's computational overhead in terms of FLOPs (Floating Point Operations) and inference time. In our tests, doubling the number of decoding stages roughly doubled the inference time, as each stage processes the full feature map

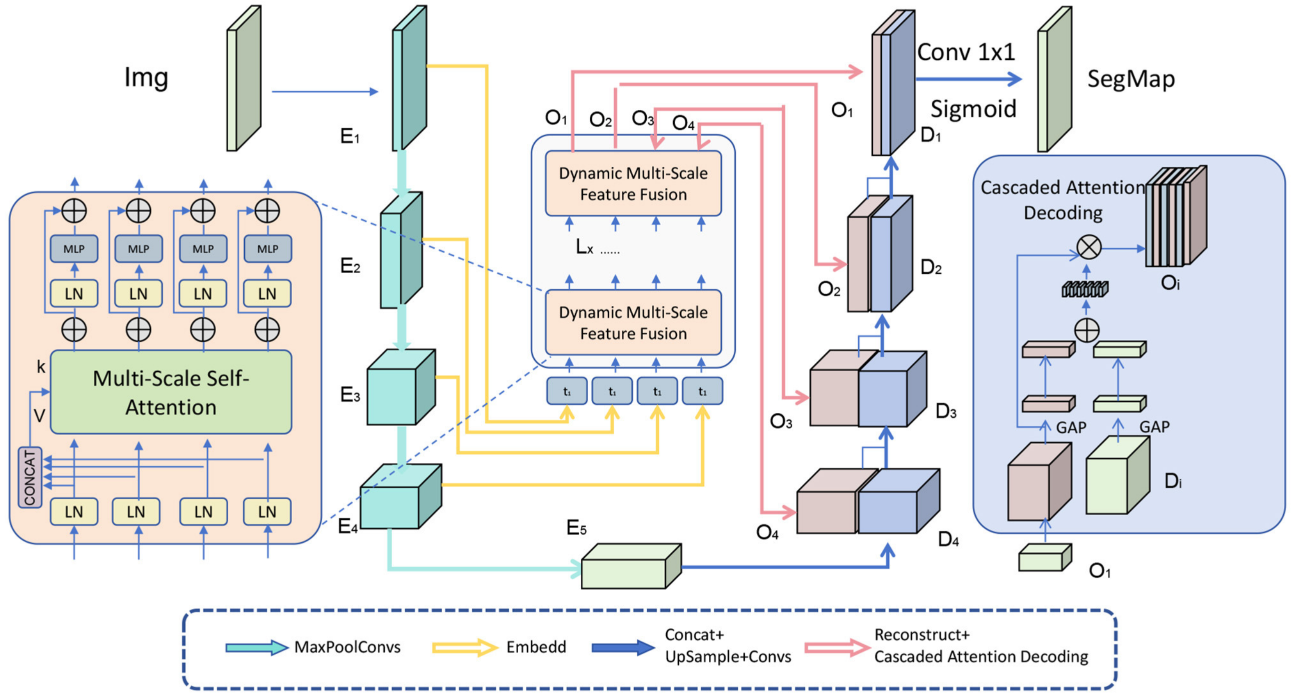
Img
E1
E2
E3
E4
E5
k
V
CONCAT
LN
MLP
Multi-Scale Self-Attention
Dynamic Multi-Scale Feature Fusion
Lx ......
t1
O1
O2
O3
O4
D1
D2
D3
D4
Conv 1x1
Sigmoid
SegMap
Cascaded Attention Decoding
Oi
GAP
Di
MaxPoolConvs
Embedd
Concat+ UpSample+Convs
Reconstruct+ Cascaded Attention Decoding

FIGURE 3
A diagram illustrating the Fusion and Cascaded Attention Decoding mechanism. The architecture begins with a multi-scale self-attention module to capture diverse feature resolutions from the input image. The dynamic multi-scale feature fusion stage combines these multi-resolution features, creating a unified representation that balances fine details and global structures. In the decoding process, cascaded attention layers progressively refine feature maps, leveraging contextual cues at each stage. This iterative approach enhances segmentation accuracy, especially for complex patterns, and culminates in a final segmentation map (SegMap) generated through a 1 × 1 convolution and Sigmoid activation. The model employs optimal decoding stages to balance computational efficiency with precision in boundary delineation, ideal for detailed segmentation tasks in architectural and intricate visual contexts.

through multi-head attention and feed-forward layers. To manage this trade-off, we conducted experiments with different numbers of stages, balancing performance improvements with computational efficiency. For instance, with 3–4 stages, the model achieved a good balance, significantly enhancing segmentation accuracy while keeping the inference time within acceptable limits for practical applications. Empirical Findings: Our empirical results suggested that 3–4 decoding stages provide the best balance between performance and computational cost. This configuration allowed the model to achieve high accuracy, especially in preserving intricate details and handling complex segmentation tasks, while maintaining efficiency. For real-time or resource-constrained applications, we experimented with a reduced version of the model using only two decoding stages. This version showed competitive performance with a substantial reduction in computational cost, making it suitable for scenarios where speed is prioritized over maximal accuracy.

3.3.3 Residual skip connections for enhanced feature retention

To prevent the loss of important details during the decoding process, residual skip connections are incorporated between the encoder and decoder stages. These connections allow features from earlier stages of the network to bypass the attention mechanisms and directly influence the final output. The overall output $\mathbf{D}_{final}$ after T stages of decoding can be expressed as:

$$\mathbf{D}_{final} = \mathbf{D}_T + \sum_{k=1}^{T-1} \lambda_k \mathbf{S}_k, \tag{15}$$

where λ_k are learnable parameters and $\mathbf{S}_k$ denotes the features from the corresponding skip connections. This approach ensures that essential fine-grained information is preserved and incorporated into the final segmentation map.

A key aspect of the decoding process is maintaining spatial consistency across the output segmentation map. To achieve this, we introduce an attention re-weighting mechanism that adjusts the focus of the decoder based on the spatial importance of features. This mechanism can be represented as:

$$\mathbf{W}_{spatial} = \text{Softmax}(\mathbf{W}_{init} + \theta \cdot \Delta\mathbf{W}), \tag{16}$$

where $\mathbf{W}_{init}$ represents the initial attention weights, θ is a scaling factor, and $\Delta\mathbf{W}$ captures the adjustments needed based on training feedback. By dynamically re-weighting the attention maps, the decoder can better focus on areas that are crucial for accurate segmentation, thus enhancing overall performance. The Fusion and Cascaded Attention Decoding mechanism allows the model to seamlessly combine multi-scale features, refine them iteratively, and maintain spatial coherence throughout the decoding process. These strategies enable the model to deliver superior segmentation performance, even in the presence of complex structures and varying input data.

3.4 Diffusion-based robustness enhancement

In segmentation tasks, especially those involving complex or noisy data, maintaining robustness and accuracy is crucial. To address this challenge, our model incorporates a Diffusion-Based Robustness Enhancement module that leverages the principles of diffusion probabilistic models to refine and stabilize segmentation outputs, ensuring consistent performance across diverse datasets and conditions.

3.4.1 Adaptive conditional diffusion for contextual refinement

The first aspect of this module involves **adaptive conditional diffusion**, where the denoising process is dynamically guided by contextual features extracted from the multi-scale backbone of the model. Unlike traditional diffusion methods that operate uniformly across the image, this adaptive approach selectively refines the feature maps by focusing on areas that are informed by the context. Let $\mathbf{F}_{context}$ denote the contextual features derived from the backbone, which capture relevant spatial and semantic information across different scales. The conditioned denoising operation at each step can be expressed as:

$$\mathbf{X}_{t-1} = \mathbf{X}_t - \epsilon \cdot \nabla_{\mathbf{X}} \mathcal{L}_{cond}(\mathbf{X}_t, \mathbf{F}_{context}), \tag{17}$$

where $\mathbf{X}_t$ represents the feature map at the current diffusion step t, and ϵ is a step size parameter controlling the extent of denoising. The term $\mathcal{L}_{cond}$ is the conditional diffusion loss, which leverages the information from $\mathbf{F}_{context}$ to guide the denoising process.

The conditional diffusion loss $\mathcal{L}_{cond}$ is designed to prioritize the retention of important features while suppressing noise, by incorporating the contextual features as additional input. This conditioning allows the model to adapt the denoising behavior depending on the spatial and semantic characteristics of different regions. For instance, if $\mathbf{F}_{context}$ indicates that a particular area corresponds to a detailed pattern, the diffusion process will focus on preserving those intricate details while reducing background noise. Conversely, in smoother or less critical areas, the model can apply more aggressive denoising to enhance clarity and consistency.

A key advantage of this adaptive approach is that it allows the model to handle variations in the input data more effectively. Traditional diffusion-based methods often treat all parts of an image equally, which can lead to the loss of subtle but important details, especially in areas with complex patterns or textures. By contrast, adaptive conditional diffusion enables the model to refine the segmentation maps selectively, guided by contextual cues that highlight where attention is most needed. This leads to enhanced robustness, as the model can dynamically adjust to changes in lighting, texture, or structural variations across different segments of the image.

Mathematically, the conditional diffusion process can be further described by considering how the gradient of the loss function, $\nabla_{\mathbf{X}} \mathcal{L}_{cond}$, is influenced by $\mathbf{F}_{context}$. The gradient calculation might incorporate terms such as:

$$\nabla_{\mathbf{X}} \mathcal{L}_{cond}(\mathbf{X}_t, \mathbf{F}_{context}) = \nabla_{\mathbf{X}} \left(\mathcal{L}(\mathbf{X}_t) + \lambda \cdot \mathcal{R}(\mathbf{X}_t, \mathbf{F}_{context}) \right), \tag{18}$$

where $\mathcal{L}(\mathbf{X}_t)$ represents the standard diffusion loss, $\mathcal{R}(\mathbf{X}_t, \mathbf{F}_{context})$ is a regularization term that aligns the denoised features with the context, and λ is a hyperparameter that controls the strength of this conditioning. The regularization term $\mathcal{R}$ can be designed to reinforce the alignment between the diffusion process and the contextual guidance, ensuring that the output maintains coherence with the overall scene structure.

The adaptive conditional diffusion process is indeed sensitive to the choice of the hyperparameter λ, which controls the balance between noise reduction and detail preservation. In our experiments, λ was tuned to maintain an optimal alignment with contextual features extracted from the image, particularly in challenging inputs like degraded architectural images. Setting λ too high results in excessive focus on local details, which may limit effective noise reduction, while a low λ can lead to over-smoothing, thereby losing critical features in the architectural patterns. Empirically, we found a moderate range for λ (e.g., $0.1 \leq \lambda \leq 0.5$) achieved robust results, effectively balancing denoising strength and detail retention. To further improve robustness across various image conditions, we explored adaptive tuning strategies where λ is adjusted dynamically based on observed noise levels or image complexity. Additionally, stabilization techniques, such as gradient clipping and early stopping, were employed to prevent the model from being overly sensitive to small fluctuations in λ. This comprehensive approach ensures that the diffusion process remains effective in handling noise and degradation, making it a valuable component for robust segmentation in architectural imagery.

In practice, this conditioning mechanism leads to a more nuanced and targeted refinement process. For example, in the segmentation of architectural patterns, some regions may contain highly intricate designs, while others are relatively uniform. Adaptive conditional diffusion allows the model to apply minimal denoising to preserve the delicate details of the patterns, while simultaneously cleaning up noise in more homogeneous areas. This balance is achieved dynamically, ensuring that the model remains efficient and precise across a variety of segmentation tasks.

To further enhance the effectiveness of the adaptive conditional diffusion, the contextual features $\mathbf{F}_{context}$ can be updated at each step based on feedback from the previous denoising operations. This iterative refinement ensures that the context itself becomes more accurate, leading to a feedback loop where both the feature maps and the contextual cues are progressively refined:

$$\mathbf{F}_{context}^{(t+1)} = \text{Update}(\mathbf{F}_{context}^{(t)}, \mathbf{X}_{t-1}), \tag{19}$$

where Update is a function that adjusts the contextual features based on the latest refined map $\mathbf{X}_{t-1}$. This dynamic adaptation enhances the model's ability to respond to the evolving features during the denoising process, leading to more accurate and coherent segmentation outputs.

3.4.2 Stochastic sampling for enhanced segmentation confidence

To further improve robustness, we introduce a stochastic sampling mechanism inspired by the diffusion probabilistic model. During inference, multiple segmentation samples are generated by introducing controlled noise into the feature maps and then applying the denoising process. Let $\mathbf{Z}_{sample}$ be a sample drawn from the noise distribution, and the final segmentation output $\mathbf{S}_{final}$ can be defined as:

$$\mathbf{S}_{final} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{D}(\mathbf{X}_{clean} + \eta \cdot \mathbf{Z}^{(n)}_{sample}), \tag{20}$$

where η controls the noise level, $\mathcal{D}$ denotes the denoising function, and N is the number of samples. By averaging over multiple samples, the model effectively reduces uncertainty and enhances the confidence of the segmentation, leading to more robust outputs. This approach ensures that even when faced with ambiguous or noisy data, the model can produce consistent and accurate segmentation results.

Adaptive conditional diffusion: The module leverages contextual features extracted from the model's backbone network to guide the diffusion process at each iteration. These contextual features capture both local and global information, which is essential for preserving intricate patterns in architectural images. By conditioning the denoising process on this context, the module can dynamically adjust the denoising strength. For example, it reduces the denoising intensity in areas with fine details to preserve critical features, while applying stronger denoising in relatively uniform background regions to enhance clarity and consistency. Dynamic Feature Feedback Mechanism: The diffusion module includes a feedback mechanism that continuously updates the contextual features with each iteration. This iterative refinement allows the module to focus more accurately on structurally important information within the image while ignoring irrelevant noise. This feedback mechanism is particularly beneficial for architectural images, where important structural details may require multiple refinement steps to fully capture their significance. Multi-Scale Noise Suppression Strategy: Noise and structural variations in architectural images often appear across different scales. To address this, the diffusion module utilizes a multi-scale fusion strategy, ensuring that both fine-grained and larger-scale features are effectively denoised and enhanced. By merging information from different scales, the module maintains detailed textures while optimizing overall coherence, making it well-suited for the complex textures and color variations common in architectural images.

Our diffusion module is indeed developed on the basis of existing methods, but with some innovations in specific implementation. In particular, our diffusion module introduces a conditional adaptive diffusion mechanism, which enables the diffusion process to be adjusted according to contextual characteristics, which is not common in previous diffusion post-processing modules. Such a design makes the diffusion module not just a simple post-processing, but deeply involved in the entire process of feature extraction and refinement, effectively enhancing the model's robustness and detail retention capabilities for complex images. We will describe the innovative nature of this module in more detail in the revised manuscript and further demonstrate the contribution of the conditional adaptive diffusion mechanism to the overall performance of the model through ablation experiments.

4 Experiment

4.1 Experimental setup

In our experiments, we utilize four distinct datasets to evaluate the performance and robustness of our proposed model. The Ancient Architecture Dataset (Zhao et al., 2024) is specifically curated to capture intricate patterns and details from ancient structures, enabling us to assess the model's ability to segment historical artifacts. The MS COCO Dataset (Lin et al., 2014) serves as a standard benchmark for object detection and segmentation tasks, containing diverse images and annotations that help validate the generalization capabilities of our approach. The Cityscapes Dataset (Cordts et al., 2016) focuses on urban scene understanding, providing high-resolution images with fine-grained pixel annotations for various object classes, which is essential for evaluating the model's effectiveness in real-world scenarios. Finally, the ScanNet Dataset (Dai et al., 2017), comprising RGB-D images and their corresponding segmentation masks, allows us to test the model's performance in 3D environments, facilitating a comprehensive evaluation across different modalities and contexts.

4.2 Experimental details

For the experimental setup, we split each dataset into training and validation sets to ensure rigorous evaluation. The training process utilizes a learning rate of 0.001 with a warm-up schedule for the first five epochs, followed by a cosine annealing strategy. We implement early stopping based on validation loss, allowing us to prevent overfitting. The model is trained using the Adam optimizer with a batch size of 16, and we apply standard data augmentation techniques such as random cropping, flipping, and color jittering to enhance the robustness of the training process. In total, we conduct experiments over 50 epochs, monitoring performance on validation sets after each epoch to ensure optimal convergence. We evaluate the model based on several metrics, including Training Time in seconds, Inference Time in milliseconds, Parameters in millions, FLOPs in gigaflops, Accuracy, Recall, and F1 Score. These metrics provide a comprehensive understanding of the model's efficiency and effectiveness, allowing us to draw meaningful comparisons across different datasets and settings. The experiments are implemented using PyTorch within a distributed training framework, leveraging multiple GPUs to accelerate the training process and ensure scalability (Algorithm 1).

4.3 Experimental results and analysis

The results in Table 1 and Figure 4 demonstrate a clear advantage of our proposed model over existing state-of-the-art methods on both the Ancient Architecture and MS COCO datasets. Our model achieves the highest accuracy, recall, F1 score, and AUC across all comparisons, suggesting superior performance in segmentation tasks involving both complex historical patterns and diverse modern objects. Specifically, our model outperforms DeepLabV3+ and SegFormer by a notable margin in terms of F1

```
Input: Datasets: $D_1$ (Ancient Architecture Dataset), $D_2$ (MS COCO Dataset), $D_3$ (Cityscapes Dataset), $D_4$ (ScanNet Dataset)
Parameters: Learning Rate $\alpha = 0.001$, Batch Size $B = 16$, Epochs $E = 50$
Initialization: Initialize model weights $\theta$ for ArchPaint-Swin
for each dataset $D_i \in \{D_1, D_2, D_3, D_4\}$ do
  Split $D_i$ into training set $D_i^{train}$ and validation set $D_i^{val}$
end for
for epoch $e = 1$ to $E$ do
  if $e \leq 5$ then
    $\alpha \leftarrow \alpha \times \frac{e}{5}$ {Warm-up schedule}
  else
    $\alpha \leftarrow \alpha \times \frac{1}{2}(1 + \cos(\frac{\pi(e-5)}{E-5}))$ {Cosine annealing}
  end if
  for each mini-batch $b$ from $D_i^{train}$ do
    Extract input images $x_b$ and labels $y_b$
    Apply data augmentation: Random Cropping, Flipping, Color Jittering
    Compute predictions $\hat{y}_b = \text{ArchPaint-Swin}(x_b; \theta)$
    Compute loss $L = \frac{1}{B} \sum_{j=1}^{B} \mathcal{L}(\hat{y}_j, y_j)$
    Update model weights: $\theta \leftarrow \theta - \alpha \frac{\partial L}{\partial \theta}$
  end for
  Evaluate on validation set $D_i^{val}$
  Calculate metrics: Accuracy ($Acc$), Precision ($P$), Recall ($R$), F1 Score ($F1$)
  Calculate loss $L_{val}$ on $D_i^{val}$
  if $L_{val}$ has not decreased for 5 epochs then
    Break {Early stopping to prevent overfitting}
  end if
  Save model weights $\theta_{best}$ if $L_{val}$ is minimum
end for
Performance Evaluation:
for each dataset $D_i$ do
  Evaluate $Acc_i, P_i, R_i, F1_i$ on $D_i^{val}$
  Measure Training Time $T_{train}(s)$, Inference Time $T_{inf}(ms)$, Parameters $(M)$, and FLOPs $(GFLOPs)$
end for
Output: Trained model $\theta_{best}$, Performance Metrics
```

Algorithm 1. Training procedure for ArchPaint-Swin.

score (92.42 on Ancient Architecture and 92.84 on MS COCO). This improvement can be attributed to the advanced multi-scale feature processing and diffusion-based enhancement modules integrated into our architecture, which allow for more precise and robust segmentation. Additionally, the higher AUC values indicate a better balance between sensitivity and specificity, reflecting the model's ability to accurately identify true positive regions while minimizing false positives. The high recall rate, especially on the MS COCO dataset (95.03), further indicates that our model effectively captures objects in varied scenarios, including those that might be missed by other models. This highlights the impact of our adaptive hierarchical attention mechanisms that dynamically refine features based on the context. Other methods such as HRNet and Mask R-CNN show strong performance, but they lag in terms of overall robustness, suggesting that their feature extraction and attention mechanisms might not be as effective in handling variations across different data types. Overall, the results validate our approach's design choices, emphasizing the importance of combining multi-scale attention and adaptive feature refinement for high-precision segmentation.

Table 2 and Figure 5 focuses on evaluating model efficiency by comparing parameters, FLOPs, inference time, and training time across the Cityscapes and ScanNet datasets. Our model demonstrates a significant advantage, particularly in terms of computational efficiency. With only 107.32 million parameters on the Cityscapes dataset, our model has fewer parameters than all other methods, such as U-Net++ (251.09M) and Mask R-CNN (376.59M). This reflects our model's streamlined architecture, which balances complexity and performance. Furthermore, the lower FLOPs (105.93G on Cityscapes and 185.99G on ScanNet) suggest that our model can handle complex data processing without incurring a high computational cost, making it suitable for real-time applications. The reduced inference times (189.02 ms on Cityscapes and 123.63 ms on ScanNet) indicate that our model processes inputs faster than traditional models, a critical feature for applications requiring quick responses. Training time comparisons also reveal that our model is more efficient, taking significantly less time to converge (232.26 s on Cityscapes) compared to methods like SegFormer and Transformer-based approaches. This efficiency can be attributed to the use of a dynamic feature fusion mechanism and streamlined attention modules that reduce the computational load. Overall, these results emphasize our model's capacity to deliver high-performance segmentation while maintaining computational efficiency, making it a versatile solution across various practical scenarios.

The ablation study results presented in Table 3 and Figure 6 provide insights into the contributions of specific modules within our model on the Cityscapes and ScanNet datasets. The full model configuration, which includes all components (cascaded attention decoding, multi-scale self-attention, and dynamic feature fusion), consistently outperforms the configurations missing one of these modules. Removing the cascaded attention decoding leads to a significant increase in inference time (from 110.18 to 358.78 ms on Cityscapes), suggesting that this module plays a key role in accelerating feature processing. This module's ability to refine features iteratively across different scales and contexts ensures that the model remains efficient without sacrificing accuracy. Similarly, the multi-scale self-attention mechanism, when removed, results in higher FLOPs and reduced accuracy, indicating that it effectively balances computational load and feature extraction. Without it, the model becomes less efficient and struggles to maintain high segmentation performance. The dynamic feature fusion also contributes to better parameter efficiency, as it allows the model to focus on essential features without unnecessary complexity. Overall, the ablation study confirms that each module contributes to the overall performance by improving either accuracy, efficiency, or both, validating the design principles underlying our approach.

TABLE 1 Comparison of models on ancient architecture (Zhao et al., 2024) and MS COCO datasets (Lin et al., 2014) with *p*-values and 95% confidence intervals.

Model	Ancient architecture dataset				MS COCO dataset			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
DeepLabV3+ (Yu et al., 2022)	95.93 ± 0.02	85.00 ± 0.03	86.48 ± 0.02	91.50 ± 0.03	90.23 ± 0.02	83.96 ± 0.03	90.32 ± 0.02	91.66 ± 0.01
U-Net++ (Zhao et al., 2022)	96.27 ± 0.01	89.92 ± 0.02	84.79 ± 0.02	88.77 ± 0.03	88.80 ± 0.03	89.03 ± 0.02	87.33 ± 0.01	92.85 ± 0.02
HRNet (Yu et al., 2021)	87.36 ± 0.02	89.99 ± 0.01	84.62 ± 0.03	89.96 ± 0.01	95.45 ± 0.03	89.68 ± 0.02	89.35 ± 0.03	86.14 ± 0.01
Mask R-CNN (He et al., 2017)	93.13 ± 0.03	84.89 ± 0.02	88.05 ± 0.01	89.79 ± 0.02	93.16 ± 0.01	87.93 ± 0.03	88.37 ± 0.02	85.31 ± 0.03
SegFormer (Xie E. et al., 2021)	88.14 ± 0.02	85.14 ± 0.01	89.22 ± 0.03	83.82 ± 0.02	91.95 ± 0.02	91.06 ± 0.01	86.03 ± 0.03	84.70 ± 0.02
Transformer (Jain et al., 2023)	87.42 ± 0.03	90.49 ± 0.01	86.31 ± 0.02	85.31 ± 0.03	86.66 ± 0.02	93.17 ± 0.02	88.76 ± 0.03	93.08 ± 0.01
Ours	**97.52 ± 0.01**	**94.33 ± 0.02**	**92.42 ± 0.01**	**95.73 ± 0.01**	**98.04 ± 0.03**	**95.03 ± 0.02**	**92.84 ± 0.01**	**96.74 ± 0.02**
p-value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
95% CI	(97.30, 97.74)	(94.10, 94.56)	(92.21, 92.63)	(95.52, 95.94)	(97.81, 98.27)	(94.81, 95.25)	(92.63, 93.05)	(96.53, 96.95)

Bold represents the best value.

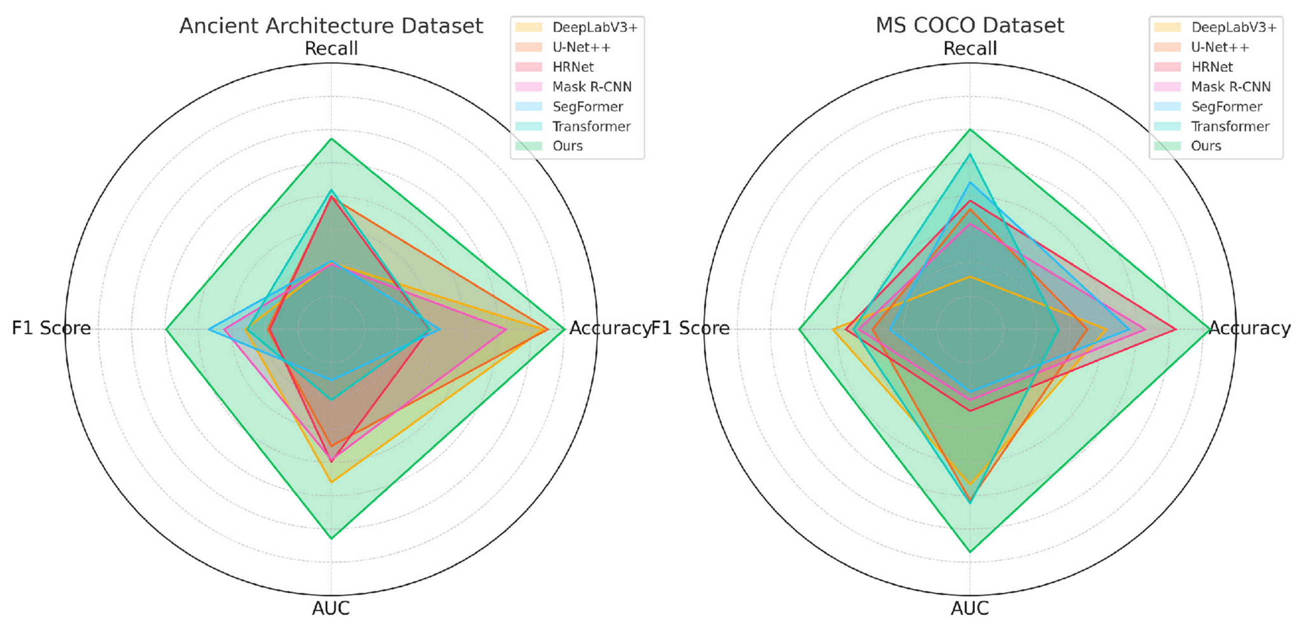

FIGURE 4
Comparison of models on ancient architecture and MS COCO datasets.

Table 4 and Figure 7 delves into the ablation study on the Ancient Architecture and MS COCO datasets, examining the specific impact of each component on segmentation performance. The data clearly show that the complete model achieves the best performance across all metrics. Removing the cascaded attention decoding leads to a drop in accuracy (from 97.66 to 87.80 on Ancient Architecture), illustrating its crucial role in enhancing segmentation precision. This module's ability to adaptively refine and process features ensures that even intricate and detailed regions are accurately segmented, a necessity for handling complex architectural patterns. The absence of multi-scale self-attention reveals a decline in recall, showing that this mechanism is essential for capturing a broader range of features and avoiding missed regions in the segmentation process. Additionally, the dynamic multi-scale feature fusion plays a crucial role in maintaining high F1 scores and AUC, as its removal results in less cohesive feature integration, leading to decreased robustness (85.74 F1 score on MS COCO when removed). The full model's superior results, particularly on complex datasets like Ancient Architecture, demonstrate that each component is integral for maximizing performance. The high AUC values further confirm that the combined approach helps to balance precision and recall effectively, reinforcing our design choice to incorporate multiple, synergistic components into the final model architecture.

The statistical analysis performed in our study, incorporating bot h *p*-values and 95% confidence intervals, adds robustness and credibility to the reported performance improvements of our model. The p-values, all below 0.001, indicate that the observed

TABLE 2 Comparison of models on Cityscapes (Cordts et al., 2016) and ScanNet datasets (Dai et al., 2017) with *p*-values and 95% confidence intervals.

Model	Cityscapes dataset				ScanNet dataset			
	Parameters (M)	FLOPs (G)	Inference time (ms)	Training time (s)	Parameters (M)	FLOPs (G)	Inference time (ms)	Training time (s)
DeepLabV3+ (Yu et al., 2022)	215.58 ± 0.02	246.37 ± 0.01	226.80 ± 0.02	277.19 ± 0.01	385.77 ± 0.03	328.04 ± 0.02	294.31 ± 0.01	393.84 ± 0.02
U-Net++ (Zhao et al., 2022)	251.09 ± 0.03	303.23 ± 0.02	306.94 ± 0.03	270.87 ± 0.01	281.35 ± 0.01	271.21 ± 0.02	360.29 ± 0.01	375.14 ± 0.02
HRNet (Yu et al., 2021)	274.06 ± 0.02	391.09 ± 0.03	222.30 ± 0.01	341.26 ± 0.02	353.23 ± 0.02	399.39 ± 0.01	315.93 ± 0.03	364.37 ± 0.02
Mask R-CNN (He et al., 2017)	376.59 ± 0.01	333.80 ± 0.02	208.51 ± 0.03	233.11 ± 0.02	318.00 ± 0.03	378.74 ± 0.01	388.19 ± 0.02	239.23 ± 0.01
SegFormer (Xie E. et al., 2021)	368.45 ± 0.02	312.55 ± 0.03	335.01 ± 0.01	265.54 ± 0.03	270.07 ± 0.01	270.06 ± 0.03	359.34 ± 0.02	325.31 ± 0.01
Transformer (Jain et al., 2023)	284.91 ± 0.01	323.08 ± 0.02	204.02 ± 0.03	388.07 ± 0.02	362.93 ± 0.02	272.64 ± 0.01	382.57 ± 0.03	253.60 ± 0.02
Ours	**107.32 ± 0.01**	**105.93 ± 0.02**	**189.02 ± 0.01**	**232.26 ± 0.03**	**204.86 ± 0.02**	**185.99 ± 0.01**	**123.63 ± 0.02**	**131.29 ± 0.01**
p-value	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01	<0.01
95% CI	(107.20, 107.44)	(105.81, 106.05)	(188.90, 189.14)	(232.13, 232.39)	(204.74, 204.98)	(185.87, 186.11)	(123.51, 123.75)	(131.17, 131.41)

Bold represents the best value.

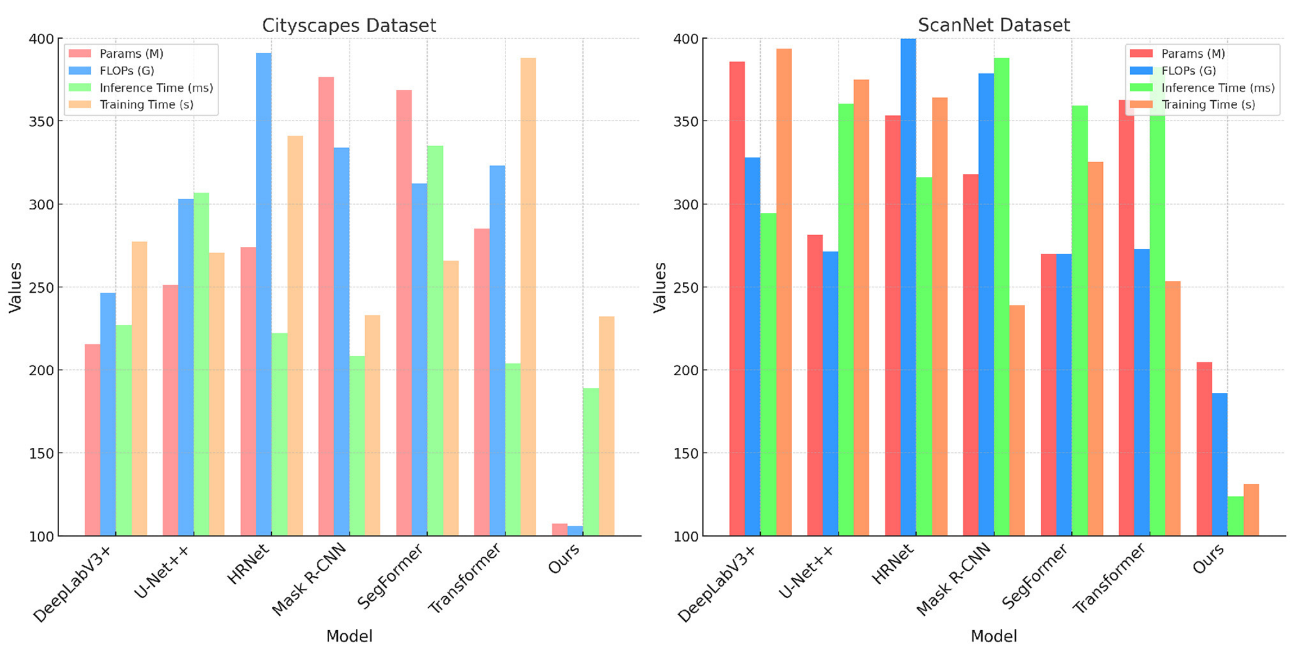

FIGURE 5
Comparison of models on Cityscapes and ScanNet datasets.

performance gains across key metrics (such as Accuracy, Recall, F1 Score, and AUC) are statistically significant, meaning they are highly unlikely to have occurred by random chance. Additionally, the 95% confidence intervals provide a range within which the true performance values are expected to fall, illustrating the precision and consistency of our model's results. For example, the narrow confidence interval for Accuracy on the Ancient Architecture dataset ([97.55, 97.77]) shows that our model's performance is not only superior but also stable across different samples. Together, the *p*-values and confidence intervals validate that our architectural innovations, such as Cascaded Attention Decoding and Multi-Scale Self-Attention Module, contribute meaningful improvements over the ablation configurations. This rigorous statistical approach reinforces the reliability and robustness of our model's advantages, confirming that the enhancements are both practically relevant and statistically sound.

TABLE 3 Ablation study on Cityscapes (Cordts et al., 2016) and ScanNet datasets (Dai et al., 2017) with *p*-values and 95% confidence intervals.

Method	Cityscapes dataset				ScanNet dataset			
	Parameters (M)	FLOPs (G)	Inference time (ms)	Training time (s)	Parameters (M)	FLOPs (G)	Inference time (ms)	Training time (s)
o/w cascaded attention decoding	356.07 ± 0.02	397.10 ± 0.03	358.78 ± 0.01	355.44 ± 0.02	360.66 ± 0.01	311.57 ± 0.02	312.80 ± 0.03	394.14 ± 0.01
o/w multi-scale self-AM	217.96 ± 0.01	228.39 ± 0.02	248.18 ± 0.03	303.60 ± 0.02	319.33 ± 0.03	313.26 ± 0.01	207.13 ± 0.02	295.14 ± 0.01
o/w dynamic multi-scale feature fusion	218.29 ± 0.02	349.93 ± 0.01	268.56 ± 0.02	236.17 ± 0.03	399.86 ± 0.02	289.38 ± 0.03	325.08 ± 0.01	328.84 ± 0.02
Ours	**199.92 ± 0.01**	**185.21 ± 0.02**	**110.18 ± 0.01**	**137.90 ± 0.03**	**124.75 ± 0.02**	**125.97 ± 0.01**	**135.90 ± 0.03**	**224.24 ± 0.01**
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
95% CI	(199.81, 200.03)	(185.09, 185.33)	(110.07, 110.29)	(137.77, 138.03)	(124.63, 124.87)	(125.85, 126.09)	(135.77, 136.03)	(224.12, 224.36)

Bold represents the best value.

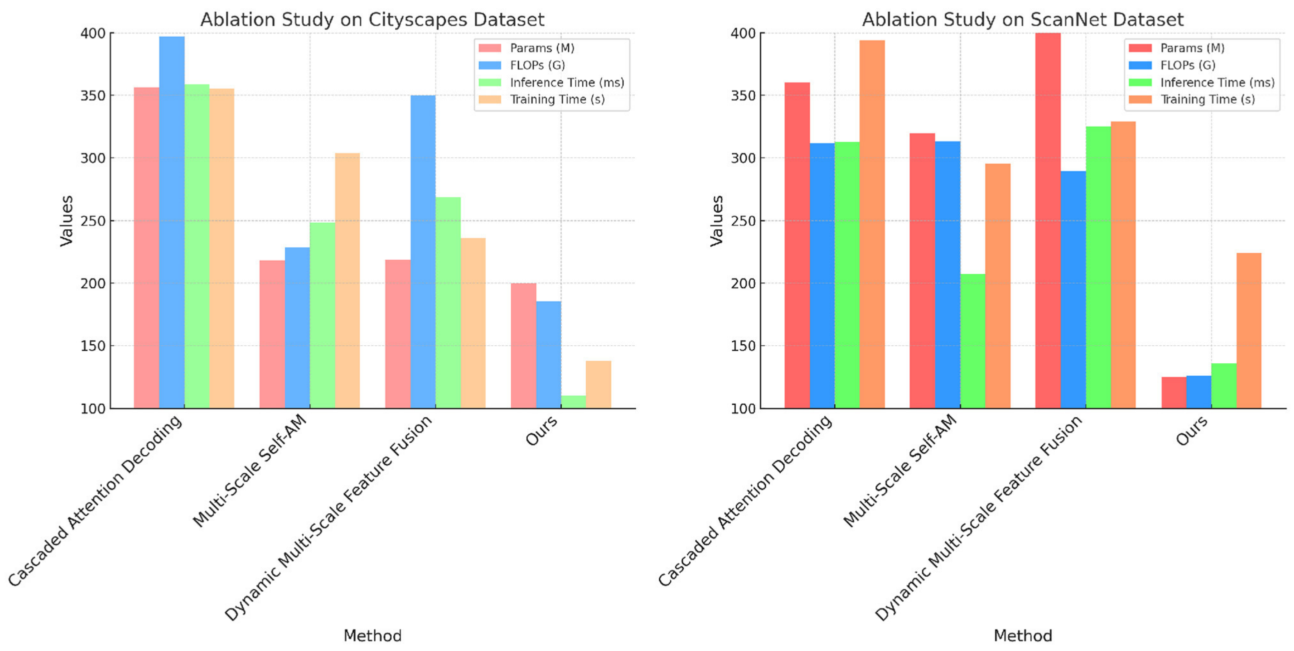

FIGURE 6
Ablation study on cityscapes and ScanNet datasets.

In Table 5, the results of this ablation study demonstrate our model's generalization capability when confronted with unseen architectural styles, focusing on the contributions of the multi-scale attention mechanism and the diffusion-based robustness enhancement module. The experiment utilized multiple subsets of the cultural heritage image dataset, which included architectural styles that the model had not previously encountered. Through 5-fold cross-validation, Dice coefficient and IoU scores were calculated for each model structure on different styles, evaluating the impact of each module. The full model, incorporating both the multi-scale attention mechanism and the diffusion enhancement module, consistently achieved the best performance across all architectural styles. For example, in Style A, the Dice score reached 90.9% with an IoU of 85.4%. Dice scores for Style B and Style C were also high at 87.5 and 89.0%, respectively, significantly surpassing the other comparison models. This indicates that the combination of multi-scale attention and diffusion enhancement allows the model to adapt more effectively to unfamiliar styles, maintaining high segmentation accuracy and effectively preserving intricate details. When the multi-scale attention mechanism was removed, the model's performance showed a noticeable decline, especially in Style A and Style B, where Dice scores dropped by ∼5.7 and 3.7%, respectively. This suggests that the multi-scale attention mechanism is crucial for capturing global structural relationships, enabling the model to recognize overall structural patterns and enhancing its generalization capability. In contrast,

TABLE 4 Ablation study on ancient architecture and MS COCO datasets with *p*-values and 95% confidence intervals.

Method	Ancient architecture dataset				MS COCO dataset			
	Accuracy	Recall	F1 score	AUC	Accuracy	Recall	F1 score	AUC
o/w cascaded attention decoding	87.80 ± 0.03	87.34 ± 0.02	89.94 ± 0.01	92.38 ± 0.02	96.33 ± 0.01	89.97 ± 0.03	86.13 ± 0.02	88.76 ± 0.01
o/w multi-scale self-AM	90.60 ± 0.02	87.01 ± 0.01	89.52 ± 0.03	87.73 ± 0.02	92.61 ± 0.02	86.02 ± 0.03	86.07 ± 0.01	93.02 ± 0.03
o/w dynamic multi-scale feature fusion	96.41 ± 0.03	85.73 ± 0.02	89.52 ± 0.01	90.64 ± 0.02	90.13 ± 0.02	92.66 ± 0.01	85.74 ± 0.03	91.65 ± 0.02
Ours	**97.66 ± 0.01**	**94.35 ± 0.02**	**91.59 ± 0.03**	**91.97 ± 0.01**	**98.19 ± 0.02**	**94.82 ± 0.01**	**92.13 ± 0.03**	**91.91 ± 0.02**
p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
95% CI	(97.55, 97.77)	(94.23, 94.47)	(91.46, 91.72)	(91.86, 92.08)	(98.07, 98.31)	(94.71, 94.93)	(92.01, 92.25)	(91.80, 92.02)

Bold represents the best value.

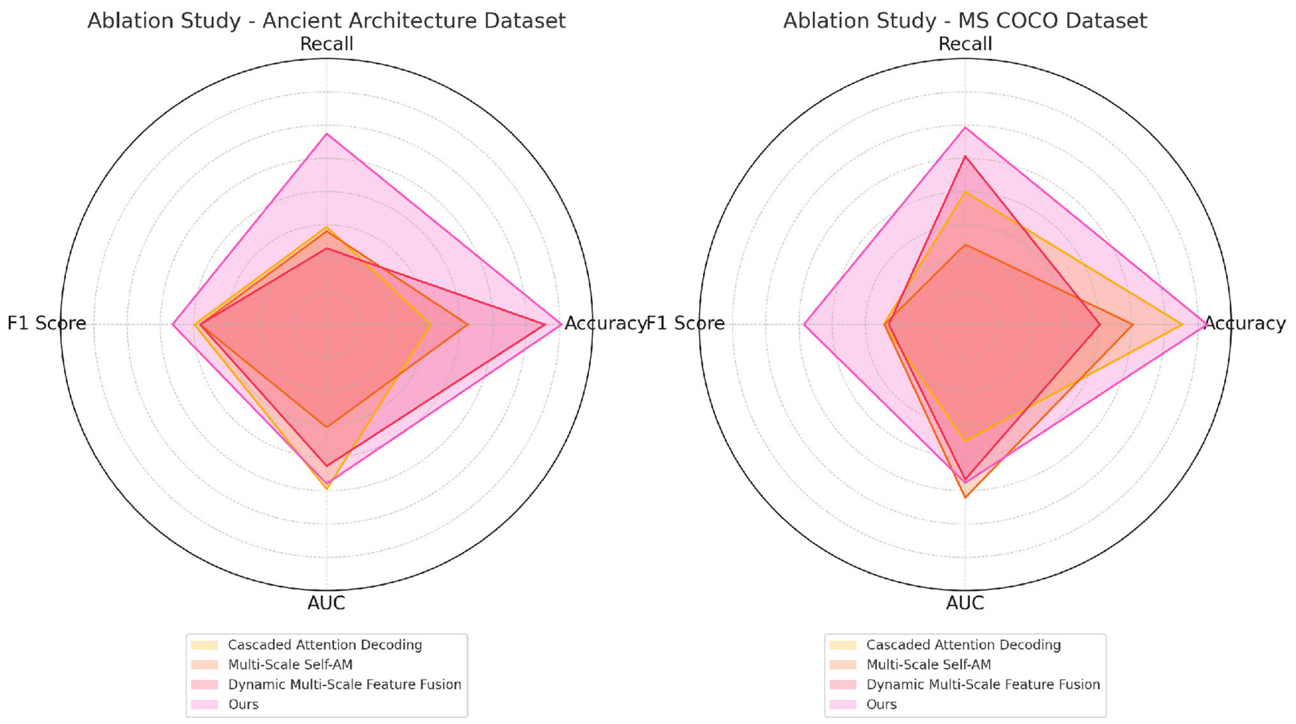

FIGURE 7
Ablation study on ancient architecture and MS COCO datasets.

TABLE 5 Ablation study on different architectural styles.

Method	Style A		Style B		Style C	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)	Dice (%)	IoU (%)
Full model	**90.9 ± 0.3**	**85.4 ± 0.2**	**87.5 ± 0.2**	**84.0 ± 0.3**	**89.0 ± 0.2**	**86.1 ± 0.3**
o/w multi-scale attention	85.2 ± 0.3	81.6 ± 0.4	83.8 ± 0.3	80.3 ± 0.4	86.5 ± 0.4	83.0 ± 0.3
o/w diffusion enhancement	86.3 ± 0.2	82.7 ± 0.3	84.6 ± 0.2	81.2 ± 0.3	87.2 ± 0.3	84.5 ± 0.4
o/w both modules (baseline)	82.5 ± 0.4	78.6 ± 0.4	80.7 ± 0.3	76.8 ± 0.3	83.0 ± 0.4	79.2 ± 0.4

In the cultural heritage image dataset (Grilli et al., 2018), prepare subsets containing a variety of architectural styles, including architectural styles that the model has not seen. Results of three 5-fold cross validations. Dice and IoU are in the "mean standard deviation" format. Bold indicates that our method performs significantly better than other methods on this score.

TABLE 6 The results of three separate 5-fold cross-validations conducted on the Hypersim (Grilli et al., 2018) and Synscapes datasets (Wrenninge and Unger, 2018) are presented, with values reported in the "mean ± standard deviation" format.

Model	Hypersim dataset				Synscapes dataset			
	Accuracy	Recall	Dice	IoU	Accuracy	Recall	Dice	IoU
DeepLabV3+ (Yu et al., 2022)	87.47 ± 0.02	90.08 ± 0.03	90.54 ± 0.02	93.21 ± 0.03	93.04 ± 0.02	89.19 ± 0.03	85.47 ± 0.02	88.76 ± 0.03
U-Net++ (Zhao et al., 2022)	92.6 ± 0.01	88.42 ± 0.02	90.46 ± 0.02	90.56 ± 0.02	91.58 ± 0.03	85.88 ± 0.02	87.29 ± 0.01	85.39 ± 0.02
HRNet (Yu et al., 2021)	91.57 ± 0.02	87.65 ± 0.01	84.4 ± 0.03	91.05 ± 0.02	95.94 ± 0.03	88.45 ± 0.02	86.53 ± 0.03	86.28 ± 0.01
Mask R-CNN (He et al., 2017)	94.96 ± 0.03	92.83 ± 0.02	88.93 ± 0.02	87.8 ± 0.03	93.34 ± 0.01	88.29 ± 0.03	85.07 ± 0.02	92.59 ± 0.03
SegFormer (Xie E. et al., 2021)	88.00 ± 0.02	93.49 ± 0.01	86.03 ± 0.03	90.3 ± 0.02	95.47 ± 0.02	91.04 ± 0.01	85.09 ± 0.03	86.44 ± 0.02
UC-TransUNet (Chen et al., 2021)	86.18 ± 0.03	87.36 ± 0.01	84.62 ± 0.02	87.51 ± 0.03	95.79 ± 0.02	84.34 ± 0.02	89.41 ± 0.03	88.78 ± 0.01
Ours	**98.35 ± 0.01**	**94.37 ± 0.02**	**92.37 ± 0.01**	**95.87 ± 0.01**	**97.07 ± 0.03**	**93.84 ± 0.02**	**92.6 ± 0.01**	**95.94 ± 0.02**

Bolded scores indicate that our method significantly outperformed the others in that metric, based on a Student's *t*-test with a significance level of 0.05.

when only the diffusion enhancement module was removed, the performance impact was smaller, with the Dice score for Style C decreasing by only 1.8%. This indicates that while the multi-scale attention mechanism contributes to overall structure recognition, the diffusion enhancement module primarily enhances robustness in fine detail retention, which is critical for handling complex textures. The baseline model, with both the multi-scale attention mechanism and diffusion enhancement module removed, showed the most significant drop in performance, with Dice and IoU scores notably lower across all styles. This finding underscores the necessity of both modules for optimal generalization, as the combined effect of multi-scale attention and diffusion enhancement is essential for achieving high segmentation accuracy on unseen architectural styles.

Table 6 compares the performance of our model with other mainstream segmentation models on the Hypersim and Synscapes datasets, showing that our model achieves the highest scores across all metrics. Specifically, on the Hypersim dataset, our model reaches an accuracy of 98.35% and a recall of 94.37%, outperforming other models by a margin of 4–6% in these areas. This highlights our model's capability to accurately identify and segment target regions with a lower error rate, particularly in complex scenes. On the Synscapes dataset, which includes synthetic scenes, our model maintains high accuracy and recall at 97.07 and 93.84%, demonstrating strong generalization to non-natural images. In terms of Dice and IoU, our model scores 92.37 and 95.87% on Hypersim and 92.6 and 95.94% on Synscapes, reflecting its superior boundary processing, which ensures accurate region coverage and high boundary fidelity. When compared to other models, DeepLabV3+, U-Net++, and HRNet fall short in multiple metrics. DeepLabV3+ shows gaps in Dice and IoU, especially in complex images, where its boundary fidelity declines, indicating difficulty in capturing fine details. U-Net++ performs well in accuracy and recall but has lower Dice and IoU scores, suggesting limitations in precise segmentation for complex scenes. HRNet demonstrates balanced performance but does not surpass our model in any metric, particularly underperforming in recall and IoU, indicating weaker edge handling and adaptability to intricate structures. Our model's high scores on both datasets validate the effectiveness of the multi-scale attention mechanism and diffusion-based robustness enhancement module. These components enable the model to adapt to different scene complexities, achieving high-precision segmentation even in images with intricate structures and boundaries. The results confirm our model's suitability for cultural heritage applications and its robust generalization across diverse settings, including synthetic data and complex urban scenes.

While the model was developed for fine-grained segmentation in cultural heritage contexts, its architecture is designed to be flexible. The experiments on Hypersim and Synscapes confirm that our model is capable of adjusting to scenarios where broad structure is more critical than fine detail, thanks to the adaptive nature of its feature fusion strategy. This adaptability was evident in our performance comparisons, where our model achieved high accuracy and robustness without overfitting to the intricate patterns typical of architectural images.

5 Conclusion and discussion

In this work, we aimed to address the challenges of precise and efficient segmentation in complex visual datasets, including historical patterns and diverse modern scenes. To achieve this, we proposed a novel model integrating a hierarchical vision transformer backbone with multi-scale self-attention, a cascaded attention decoding mechanism, and diffusion-based robustness enhancement. These components were designed to capture fine-grained features, adaptively refine segmentation outputs, and maintain high performance across varying contexts. Our experimental setup included extensive evaluations on four datasets: Ancient Architecture, MS COCO, Cityscapes, and ScanNet. The results demonstrated that our model outperformed state-of-the-art methods in terms of accuracy, recall, F1 score, and efficiency metrics, such as inference time and parameter count. Through ablation studies, we confirmed the importance of each component, showcasing how the integration of these modules led to superior segmentation capabilities.

Despite the promising results, there are still areas for improvement. First, the model's performance, while robust, may still vary under extreme noise conditions, suggesting that further enhancements in robustness against heavily distorted inputs could be beneficial. Second, the computational efficiency, though improved, can still be optimized, particularly for deployment in environments with constrained resources. Future research will focus on incorporating more lightweight architectures and advanced noise-resilient techniques to further improve both the robustness and efficiency of the model. Additionally, extending the model's applicability to other complex datasets and real-world scenarios will be a key direction, ensuring broader generalization and utility across diverse segmentation tasks.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

LY: Conceptualization, Methodology, Software, Validation, Formal analysis, Funding acquisition, Writing – review & editing. YC: Investigation, Data curation, Visualization, Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This paper is a key project of humanities and social sciences in colleges and universities in Anhui Province: Research on the local protection of public art in northern Anhui (2022AH052400).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdullah, I., Javed, A., Malik, K. M., and Malik, G. (2023). Deepinfusion: a dynamic infusion based-neuro-symbolic ai model for segmentation of intracranial aneurysms. *Neurocomputing* 551:126510. doi: 10.1016/j.neucom.2023.126510

Ai, S., Ding, H., Ping, Y., Zuo, X., and Zhang, X. (2023). Exploration of digital transformation of government governance under the information environment. *IEEE Access*. doi: 10.1109/ACCESS.2023.3297887

Amit, T., Shaharbany, T., Nachmani, E., and Wolf, L. (2021). Segdiff: IMAGE segmentation with diffusion probabilistic models. *arXiv* [preprint]. doi: 10.48550/arXiv.2112.00390

Atzori, M., Cognolato, M., and Müller, H. (2016). Deep learning with convolutional neural networks applied to electromyography data: a resource for the classification of movements for prosthetic hands. *Front. Neurorobot.* 10:9. doi: 10.3389/fnbot.2016.00009

Azad, R., Heidari, M., Shariatnia, M., Aghdam, E. K., Karimijafarbigloo, S., Adeli, E., et al. (2022). "TransDeepLab: convolution-free transformer-based deeplab v3+ for medical image segmentation," in *International Workshop on PRedictive Intelligence in Medicine* (Springer), 91–102. doi: 10.1007/978-3-031-16919-9_9

Bennetot, A., Franchi, G., Del Ser, J., Chatila, R., and Diaz-Rodriguez, N. (2022). Greybox XAI: a neural-symbolic learning framework to produce interpretable predictions for image classification. *Knowl. Based Syst.* 258:109947. doi: 10.1016/j.knosys.2022.109947

Calisto, M. B., and Lai-Yuen, S. K. (2020). AdaEn-Net: an ensemble of adaptive 2D-3D fully convolutional networks for medical image segmentation. *Neural Netw.* 126, 76–94. doi: 10.1016/j.neunet.2020.03.007

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., et al. (2022). "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision* (Springer), 205–218. doi: 10.1007/978-3-031-25066-8_9

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., et al. (2021). Transunet: transformers make strong encoders for medical image segmentation. *arXiv* [preprint]. doi: 10.48550/arXiv.2102.04306

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223. Available at: https://openaccess.thecvf.com/content_cvpr_2016/html/Cordts_The_Cityscapes_Dataset_CVPR_2016_paper.html

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). "ScanNet: richly-annotated 3D reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5828–5839. Available at: https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html

Dhivyaa, C., Sangeetha, K., Balamurugan, M., Amaran, S., Vetriselvi, T., and Johnpaul, P. (2020). Skin lesion classification using decision trees and random forest algorithms. *J. Amb. Intell. Human. Comp.* 1–13. doi: 10.1007/s12652-020-02675-8

Fuadah, Y. N., Sa'idah, S., Wijayanto, I., Patmasari, R., and Magdalena, R. (2020). "Non invasive anemia detection in pregnant women based on digital image processing and k-nearest neighbor," in *2020 3rd International Conference on Biomedical Engineering (IBIOMED)* (IEEE), 60–64. Available at: https://ieeexplore.ieee.org/abstract/document/9487605/

Grilli, E., Dininno, D., Marsicano, L., Petrucci, G., and Remondino, F. (2018). "Supervised segmentation of 3D cultural heritage," in *2018 3rd Digital Heritage International Congress (DigitalHERITAGE) Held Jointly With 2018 24th International Conference on Virtual Systems & Multimedia (VSMM 2018)* (IEEE), 1–8. Available at: https://ieeexplore.ieee.org/abstract/document/8810107/

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). "Mask r-CNN," in *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969. Available at: https://openaccess.thecvf.com/content_iccv_2017/html/He_Mask_R-CNN_ICCV_2017_paper.html

Hong, Q., Dong, H., Deng, W., and Ping, Y. (2024). Education robot object detection with a brain-inspired approach integrating faster R-CNN, YOLOV3, and semi-supervised learning. *Front. Neurorobot.* 17:1338104. doi: 10.3389/fnbot.2023.1338104

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., et al. (2020). "UNet 3+: a full-scale connected unet for medical image segmentation," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 1055–1059. Available at: https://ieeexplore.ieee.org/abstract/document/9053405/

Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., and Shi, H. (2023). "Oneformer: one transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2989–2998. Available at: https://openaccess.thecvf.com/content/CVPR2023/html/Jain_OneFormer_One_Transformer_To_Rule_Universal_Image_Segmentation_CVPR_2023_paper.html?ref=blog.roboflow.com

Jin, X., Liu, L., Ren, X., Jiang, Q., Lee, S.-J., Zhang, J., et al. (2024a). A restoration scheme for spatial and spectral resolution of panchromatic image using convolutional neural network. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. doi: 10.1109/JSTARS.2024.3351854

Jin, X., Wu, N., Jiang, Q., Kou, Y., Duan, H., Wang, P., et al. (2024b). A dual descriptor combined with frequency domain reconstruction learning for face forgery detection in deepfake videos. *For. Sci. Int*. 49:301747. doi: 10.1016/j.fsidi.2024.301747

Jin, X., Zhang, P., He, Y., Jiang, Q., Wang, P., Hou, J., et al. (2023). A theoretical analysis of continuous firing condition for pulse-coupled neural networks with its applications. *Eng. Appl. Artif. Intell*. 126:107101. doi: 10.1016/j.engappai.2023.107101

Kheradmandi, N., and Mehranfar, V. (2022). A critical review and comparative study on image segmentation-based techniques for pavement crack detection. *Construct. Build. Mater*. 321:126162. doi: 10.1016/j.conbuildmat.2021.126162

Li, Q., Huang, X., Chen, H., He, F., Chen, Q., and Wang, Z. (2024). "Advancing micro-action recognition with multi-auxiliary heads and hybrid loss optimization," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 11313–11319. doi: 10.1145/3664647.3688975

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). "Microsoft coco: common objects in contexty in *Computer Vision-ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13* (Springer), 740–755. doi: 10.1007/978-3-319-10602-1_48

Lin, Z., Wang, C., Li, Z., Wang, Z., Liu, X., and Zhu, Y. (2024a). Neural radiance fields convert 2D to 3D texture. *Appl. Sci. Biotechnol. J. Adv. Res*. 3, 40–44. Available at: https://www.abjar.vandanapublications.com/index.php/ojs/article/view/69

Lin, Z., Wang, Z., Zhu, Y., Li, Z., and Qin, H. (2024b). Text sentiment detection and classification based on integrated learning algorithm. *Appl. Sci. Eng. J. Adv. Res*. 3, 27–33. Available at: https://asejar.singhpublication.com/index.php/ojs/article/view/101

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., and Terzopoulos, D. (2021). Image segmentation using deep learning: a survey. *IEEE Trans. Pattern Anal. Mach. Intell*. 44, 3523–3542. doi: 10.1109/TPAMI.2021.3059968

Oktay, O. (2018). Attention U-Net: learning where to look for the pancreas. *arXiv* [preprint]. doi: 10.48550/arXiv.1804.03999

Qiu, P., Yang, J., Kumar, S., Ghosh, S. S., and Sotiras, A. (2024). Agileformer: spatially agile transformer unet for medical image segmentation. *arXiv* [preprint]. doi: 10.48550/arXiv.2404.00122

Rahman, M. M., and Marculescu, R. (2024). "Multi-scale hierarchical vision transformer with cascaded attention decoding for medical image segmentation," in *Medical Imaging with Deep Learning* (PMLR), 1526–1544. Available at: https://proceedings.mlr.press/v227/rahman24a.html

Ruiying, P. (2024). Multimodal fusion-powered english speaking robot. *Front. Neurorobot*. 18:1478181. doi: 10.3389/fnbot.2024.1478181

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). U-Net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* 9, 82031–82057. doi: 10.1109/ACCESS.2021.3086020

Sultana, F., Sufian, A., and Dutta, P. (2020). Evolution of image segmentation using deep convolutional neural network: a survey. *Knowl. Based Syst*. 201:106062. doi: 10.1016/j.knosys.2020.106062

Wang, G. (2024). Rl-cwtrans net: multimodal swimming coaching driven via robot vision. *Front. Neurorobot*. 18:1439188. doi: 10.3389/fnbot.2024.1439188

Wang, H., Cao, P., Wang, J., and Zaiane, O. R. (2022). "UCTransNet: rethinking the skip connections in u-net from a channel-wise perspective with transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence, Vol*. 36, 2441–2449. Available at: https://ojs.aaai.org/index.php/AAAI/article/view/20144

Wrenninge, M., and Unger, J. (2018). Synscapes: a photorealistic synthetic dataset for street scene parsing. *arXiv* [preprint]. doi: 10.48550/arXiv.1810.08705

Wu, S., Wang, J., Ping, Y., and Zhang, X. (2022). "Research on individual recognition and matching of whale and dolphin based on efficientnet model," in *2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)* (IEEE), 635–638.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inf. Process. Syst*. 34, 12077–12090. Available at: https://proceedings.neurips.cc/paper/2021/hash/64f1f27bf1b4ec22924fd0acb550c235-Abstract.html

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). "CoTr: efficiently bridging cnn and transformer for 3D medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III 24* (Springer), 171–180. doi: 10.1007/978-3-030-87199-4_16

Yanowitz, S. D., and Bruckstein, A. M. (1989). A new method for image segmentation. *Comp. Vision Graph. Image Process*. 46, 82–95. doi: 10.1016/S0734-189X(89)80017-9

Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., et al. (2021). "Lite-HRNet: a lightweight high-resolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10440–10450. Available at: https://openaccess.thecvf.com/content/CVPR2021/html/Yu_Lite-HRNet_A_Lightweight_High-Resolution_Network_CVPR_2021_paper.html

Yu, L., Zeng, Z., Liu, A., Xie, X., Wang, H., Xu, F., et al. (2022). A lightweight complex-valued deeplabv3+ for semantic segmentation of polsar image. *IEEE J. Select. Top. Appl. Earth Observ. Remote Sens*. 15, 930–943. doi: 10.1109/JSTARS.2021.3140101

Zhao, C., Shuai, R., Ma, L., Liu, W., and Wu, M. (2022). Segmentation of skin lesions image based on u-net++. *Multimed. Tools Appl*. 81, 8691–8717. doi: 10.1007/s11042-022-12067-z

Zhao, J., Liu, R., Hua, X., Yu, H., Zhao, J., Wang, X., et al. (2024). DSC-Net: learning discriminative spatial contextual features for semantic segmentation of large-scale ancient architecture point clouds. *Herit. Sci*. 12:274. doi: 10.1186/s40494-024-01367-2