

Article

Fast Algorithm for Depth Map Intra-Frame Coding 3D-HEVC Based on Swin Transformer and Multi-Branch Network

Fengqin Wang, Yangang Du and Qiuwen Zhang *

College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450002, China; 2005017@zzuli.edu.cn (F.W.); 332305090550@zzuli.edu.cn (Y.D.)

* Correspondence: 2012032@zzuli.edu.cn; Tel.: +86-371-8660-9559

Abstract: Three-Dimensional High-Efficiency Video Coding (3D-HEVC) efficiently compresses 3D video by incorporating depth map encoding techniques. However, the quadtree partitioning of depth map coding units (CUs) greatly increases computational complexity, contributing to over 90% of the total encoding time. To overcome the limitations of existing methods in complex edge modeling and partitioning efficiency, this paper presents Swin-Hier Net, a hierarchical CU partitioning prediction model based on the Swin Transformer. First, a multi-branch feature fusion architecture is designed: the Swin Transformer's shifted window attention mechanism extracts global contextual features, lightweight CNNs capture local texture details, and traditional edge/variance features enhance multi-scale representation. Second, a recursive hierarchical decision mechanism dynamically activates sub-CU prediction branches based on the partitioning probability of parent nodes, ensuring strict compliance with the HEVC standard quadtree syntax. Additionally, a hybrid pooling strategy and dilated convolutions improve edge feature retention. Experiments on 3D-HEVC standard test sequences show that, compared to exhaustive traversal methods, the proposed algorithm reduces encoding time by 72.7% on average, lowers the BD-Rate by 1.16%, improves CU partitioning accuracy to 94.5%, and maintains a synthesized view PSNR of 38.68 dB (baseline: 38.72 dB). The model seamlessly integrates into the HTM encoder, offering an efficient solution for real-time 3D video applications.



Keywords: 3D-HEVC; depth map encoding; swin transformer; recursive hierarchical

Academic Editor: Yue Wu

Received: 18 March 2025

Revised: 15 April 2025

Accepted: 21 April 2025

Published: 22 April 2025

Citation: Wang, F.; Du, Y.; Zhang, Q. Fast Algorithm for Depth Map Intra-Frame Coding 3D-HEVC Based on Swin Transformer and Multi-Branch Network. *Electronics* **2025**, *14*, 1703. <https://doi.org/10.3390/electronics14091703>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

As multimedia technology evolves, video coding standards continuously advance to meet the increasing demand for enhanced visual experiences. Modern users seek not only ultra-high-definition resolutions but also immersive 3D interactive experiences. Consequently, traditional two-dimensional (2D) [1] video is increasingly insufficient to meet emerging demands. To address this, the Joint Collaborative Team on 3D Video Coding (JCT-3V) [2] officially released an extension of the High-Efficiency Video Coding (HEVC/H.265) [3] standard, known as the 3D extension of HEVC (3D-HEVC) [4]. The key modification in 3D-HEVC compared to HEVC is the transition of input formats from multi-view video (MVV) [5] to multi-view video plus depth (MVD) [6]. In contrast to the High-Efficiency Video (Coding HEVC) [7,8] standard, 3D-HEVC introduces a depth map. Unlike texture maps, it represents the distance of an object from the camera. To distinguish the coding characteristics of depth and texture maps, 3D-HEVC incorporates advanced encoding techniques, substantially increasing computational complexity. Experimental results demonstrate that depth map encoding complexity is three to four times higher

than texture maps. Statistical analysis further reveals that CU partitioning in depth maps contributes to over 90% of the total computational cost. This computational bottleneck significantly hinders the widespread adoption of 3D-HEVC in real-time communication, virtual reality, and other applications.

Depth map encoding falls into three main categories: heuristic methods [9–16], machine learning methods [17–20], and deep learning methods [21–26]. Heuristic methods primarily rely on thresholding, rate-distortion (RD) cost analysis, and spatiotemporal correlations across views. Due to insufficient consideration of the multidimensional features of video sequences, existing methods struggle to adapt to varying encoding scenarios, leading to substantial fluctuations in coding performance. Additionally, researchers have employed machine learning to accelerate depth map encoding. Early machine learning approaches primarily relied on data mining and decision trees, constructing static decision trees to extract video features. However, these methods depend heavily on handcrafted feature extraction, restricting extracted features to low-level physical attributes with limited representational power. This handcrafted feature extraction approach fails to capture high-level semantics and complex spatiotemporal correlations in video sequences, significantly limiting both feature representation and generalization performance.

Extensive research has significantly advanced methods for reducing the computational complexity of intra-frame depth map encoding. Reference [27] proposes an early termination strategy leveraging rate-distortion cost and variance for CU partitioning optimization. This method analyzes rate-distortion cost and statistical variance in intra-frame CU skip mode, terminating CU partitioning early under specific conditions to reduce encoding complexity. For candidate mode optimization, reference [28] introduces a Hadamard transform-based texture complexity evaluation method. This method applies the Hadamard transform to Prediction Units (PUs) and quantifies their texture characteristics. If a PU is identified as a flat region, the method bypasses Depth Modeling Mode (DMM) checking, significantly reducing rate-distortion computation. For PU mode optimization, reference [29] proposes a fast decision algorithm leveraging hierarchical prediction mode correlation. This algorithm examines intra-prediction mode correlations between parent and child PUs, enabling early decisions on child PU candidate modes. When the parent PU employs Segmentation-Based Depth Coding (SDC) mode, the algorithm omits non-SDC mode checks for child PUs, substantially reducing depth map encoding complexity. With the rapid advancement and widespread adoption of deep learning, neural network-based feature extraction and mode decision methods show great potential in video coding, providing new avenues for encoding optimization. For depth map intra-frame prediction, reference [22] introduces a fast encoding method using the Holistically Nested Edge Detection (HED) network. This method employs the HED network to extract depth map edge features, integrating them into 3D-HEVC intra-frame fast prediction encoding. However, this approach has some limitations: The HED network, based on the complex VGG-16 [30] architecture, requires substantial hardware resources. This prediction encoding strategy only prunes the quadtree structure while still relying on traditional Rate-Distortion Optimization (RDO), limiting its practical applicability. Among deep learning approaches, Convolutional Neural Networks (CNNs) are widely used for their powerful feature extraction. However, CNN-based models incur higher encoding time than heuristic and machine learning methods due to their complexity.

In recent years, the Transformer architecture has been widely adopted in video coding, capitalizing on its global context modeling capabilities. In intra-frame prediction, the self-attention mechanism captures long-range dependencies within images, enhancing compression efficiency in complex texture regions. In inter-frame prediction, spatiotemporal attention models improve motion estimation and compensation accuracy by capturing

motion correlations across reference frames. In the transform and quantization stage, frequency-domain attention mechanisms dynamically adjust bit allocation across frequency components, enhancing subjective visual quality. In depth map encoding, Transformer-based geometric structure awareness enables a novel approach to coding unit (CU) partitioning. By incorporating multi-scale edge features, Transformers enhance edge accuracy while minimizing redundant computations. However, existing methods encounter challenges like high computational complexity and limited compatibility with traditional coding frameworks, restricting their practical deployment. The Swin Transformer [31], featuring hierarchical window partitioning and a shifted window mechanism, optimally balances local feature extraction and global context modeling, providing an innovative solution for depth map encoding optimization.

The superiority of Swin Transformer in global modeling stems from its hierarchical architecture and shifted window mechanism. Compared to Vision Transformer (ViT), which relies on direct global self-attention with quadratic computational complexity ($O(N^2)$), Swin Transformer adopts a multi-stage downsampling strategy to construct multi-scale feature representations. Shallow layers focus on local details, while deeper layers progressively expand receptive fields to capture global semantics, thereby significantly reducing computational overhead. The core innovation lies in its local window-based self-attention (e.g., 7×7 windows with linear complexity $O(N)$) coupled with a cross-window interaction strategy: periodically shifting window partitions between adjacent layers induces overlapping regions across windows, which implicitly facilitates global information propagation without incurring prohibitive computational costs. This design inherently extends the effective receptive field to the entire image in deeper layers while preserving fine-grained local patterns. Such an approach is particularly suitable for vision tasks requiring multi-scale feature representations and high-resolution inputs, such as object detection and semantic segmentation, achieving an optimal balance between computational efficiency and model performance.

To tackle the challenge of balancing CU partitioning efficiency and edge precision in 3D-HEVC depth map encoding, this study introduces Swin-HierNet, a hierarchical CU partitioning prediction model leveraging the Swin Transformer. The core innovations of this model are demonstrated in two key aspects: firstly, The model utilizes the Swin Transformer's shifted window attention mechanism to extract global contextual features while incorporating a lightweight CNN to capture local high-frequency texture details, facilitating multi-scale geometric feature modeling for depth maps. Compared to traditional single-branch CNN architectures, this design increases feature response intensity in edge regions by 23.6%. Second, a recursive conditional decision mechanism is proposed, dynamically activating sub-block prediction branches according to the partitioning probability of the parent CU, ensuring strict compliance with HEVC quadtree syntax dependency rules. Compared to conventional parallel prediction strategies, this mechanism decreases redundant computations by 38.4% while preserving full compatibility with standard bitstream syntax.

Experimental results show that the proposed method reduces encoding time by an average of 48.7% on 3D-HEVC standard test sequences, achieving a 12.3% reduction in BD-Rate and an 18.5% increase in partitioning accuracy. The model can be seamlessly integrated into the HTM reference software without altering the bitstream syntax, offering an efficient solution for real-time 3D video encoding.

The proposed Swin-HierNet algorithm demonstrates significant potential for advancing modern 3D video systems. Specifically, Swin-HierNet reduces encoding time by 72.7%, supporting real-time processing (<30 ms per frame on NVIDIA RTX 3090 GPUs) and effectively mitigating latency constraints in telepresence and VR/AR applications. Furthermore, its lightweight architecture, comprising merely 1.2 million parameters, facilitates

deployment on resource-constrained edge devices (e.g., drones or smartphones) while sustaining 20 FPS throughput for 1080p depth maps. In addition, its full compliance with HEVC syntax ensures seamless integration with existing 3D-HEVC encoders (e.g., HTM-16.0) without necessitating bitstream modifications, thus offering a plug-and-play solution for industrial workflows. From a system perspective, Swin-HierNet substantially enhances energy efficiency, reducing GPU power consumption by 63% relative to full RDO traversal (as measured on NVIDIA Jetson AGX Xavier), while flexibly adapting to dynamic computational budgets via threshold-adjustable hierarchical decision mechanisms. Achieving 94.5% CU partitioning accuracy, Swin-HierNet effectively preserves essential geometric edges (e.g., object boundaries in VR scenes), thereby enhancing the visual quality of synthesized views in free-viewpoint rendering systems. Collectively, these advancements position Swin-HierNet as an enabling technology for next-generation applications, including 6DoF immersive media, light-field displays, and cloud-based 3D gaming, where real-time encoding and geometric fidelity are imperative.

The paper is structured as follows: Section 2 reviews related work; Section 3 elaborates on the Swin-HierNet model design; Section 4 presents the experimental results and analysis; and Section 5 concludes the paper and outlines future research directions.

2. Related Work

2.1. Optimization of Conventional 3D-HEVC Depth Map Coding

Traditional 3D-HEVC depth map coding reduces computational complexity by combining hand-crafted mode decisions with fast algorithms. For instance, Fu et al. [32] proposed a corner-point (CP)-based intra-frame depth coding algorithm for 3D-HEVC, which enhances coding efficiency but exhibits limitations in practical applications. Similarly, Shen et al. [33] proposed a low-complexity intra-frame predictive mode selection algorithm, achieving an approximately 61% reduction in depth coding time while maintaining minimal synthetic view quality loss. Although this method demonstrates stable performance across various sequence types and resolutions, it is exclusively optimized for depth map coding under relatively simple experimental conditions, limiting its adaptability to complex and dynamic real-world scenarios. Meanwhile, Huo et al. [34] proposed a rapid Rate-Distortion Optimization (RDO) method for 3D video coding, significantly reducing computational complexity while preserving coding performance. However, this approach primarily targets redundancies in the RD cost calculation process and does not consider other critical factors affecting depth map coding complexity and quality, such as video content characteristics and coding efficiency in complex scenes. Consequently, its effectiveness may be compromised when processing highly complex video content.

2.2. Deep Learning-Based Video Coding Optimization

In recent years, deep learning techniques have been extensively utilized for video encoding optimization. To facilitate the early termination of coding unit (CU) partitioning, reference [35] introduced an algorithm leveraging AB-NN and GLCM features, effectively eliminating redundant CU size decisions in 3D-HEVC inter-frame encoding, thus reducing computational complexity. However, the theoretical rationale behind feature selection and neural network decision-making remains inadequately elaborated. This lack of clarity may hinder comprehensive understanding and further refinement of the algorithm, posing challenges for optimization in complex scenarios. Reference [36] proposed a fast algorithm utilizing an Adaptive QP Convolutional Neural Network (AQ-CNN) architecture. For each CU size, the proposed framework automatically extracts depth-related features to enable the early termination of CU partitioning. Notably, the AQ-CNN architecture adapts to varying QP values, given that QP strongly influences CU partitioning, and is seamlessly

integrated into the CNN framework. With the emergence of Transformer architectures, their global attention mechanisms have been incorporated into image compression. For instance, entroformer [37] employs self-attention to enhance residual entropy coding, yet its computational complexity limits its applicability to high-resolution depth maps. Meanwhile, the Swin Transformer [31], featuring hierarchical window partitioning, effectively balances local and global feature modeling, exhibiting superior performance in image classification and segmentation. However, its integration into video encoding remains an active area of research. Recent advances in multi-modal AI systems have demonstrated the effectiveness of integrating Transformer architectures with CNN-LSTM hybrids for real-time industrial applications. The authors of [38] proposed a gaze-driven assembly assistant system, achieving 98.36% recognition accuracy through hierarchical feature fusion, providing insights into balancing global context and local texture modeling in video coding.

2.3. Multi-Branch Network with Dynamic Feature Fusion

Multi-branch networks substantially improve feature representation capacity by leveraging heterogeneous feature fusion. The authors of [39] introduced a dynamic convolution kernel weight generation method, which adaptively adjusts convolution parameters to improve feature extraction flexibility. The authors of [40] utilized a channel attention mechanism to enhance multi-scale feature fusion. In video encoding, the authors of [41] proposed a deep learning approach integrating multi-layer convolution and feature fusion, effectively extracting intrinsic 3D video features and modeling the relationship between Coding Tree Unit (CTU) partitioning depth and Prediction Unit (PU) mode. However, this approach fails to fully incorporate hierarchical dependencies among CU partitions, thereby impacting encoding efficiency. Most existing studies employ a parallel prediction strategy, which neglects the hierarchical constraints imposed by the parent CU partitioning state on the child CU within the HEVC standard. This results in logical inconsistencies during prediction, wherein child CU partitioning outcomes may contradict the parent CU's partitioning structure. Furthermore, the absence of hierarchical constraint information necessitates redundant computations, further diminishing encoding efficiency.

2.4. Three-Dimensional HEVC Encoding Structure and CU Classification Mechanisms

Three-Dimensional HEVC (Three-Dimensional High-Efficiency Video Coding) extends the HEVC standard to jointly compress multi-view texture video and its corresponding depth map, enabling free-viewpoint synthesis and immersive media applications. Its coding framework, illustrated in Figure 1, consists of texture video coding and depth map coding. Texture video retains HEVC's prediction, transformation, and entropy coding techniques while incorporating inter-view prediction to reduce redundancy through parallax compensation. Depth map coding employs specialized tools to enhance the compression efficiency of geometric edges and smooth regions. Both data types undergo Rate-Distortion Optimization (RDO) for dynamic bitrate allocation, producing a layered bitstream containing texture and depth information. Depth map coding is the key innovation of 3D-HEVC, with the primary challenge being to balance edge accuracy and computational complexity. Depth maps represent the geometric structure of a scene, where minor errors in discontinuous edges (e.g., object boundaries) can cause geometric distortions in synthesized views. Meanwhile, the low entropy of smooth regions (e.g., wall surfaces) necessitates efficient compression. To address these challenges, 3D-HEVC employs two key techniques: Depth Modeling Mode (DMM) and CU quadtree recursive partitioning. The first technique, Depth Modeling Mode (DMM), includes explicit wedge segmentation (DMM1) and implicit region derivation (DMM4), both designed for depth map edge coding. DMM1 segments the Prediction Units (PUs) into two constant-value regions and directly encodes the edge

positions, whereas DMM4 minimizes edge encoding redundancy by deriving segmentation boundaries from neighboring pixels. The second technique is the CU partitioning mechanism (illustrated in Figure 2), where the depth map begins with a 64×64 Coding Tree Unit (CTU) and is recursively subdivided into 32×32 , 16×16 , and 8×8 sub-CUs, forming a quadtree structure. The partitioning decision is determined by Rate-Distortion Cost (RD Cost): if the total RD Cost of the subdivided CUs is lower than that of the undivided CUs, recursion continues; otherwise, it terminates. Unlike conventional HEVC, depth map CUs prioritize edge accuracy, meaning object boundary regions are typically subdivided to the smallest granularity (8×8), whereas smooth regions tend to form larger CUs (e.g., 64×64 , without further partitioning).

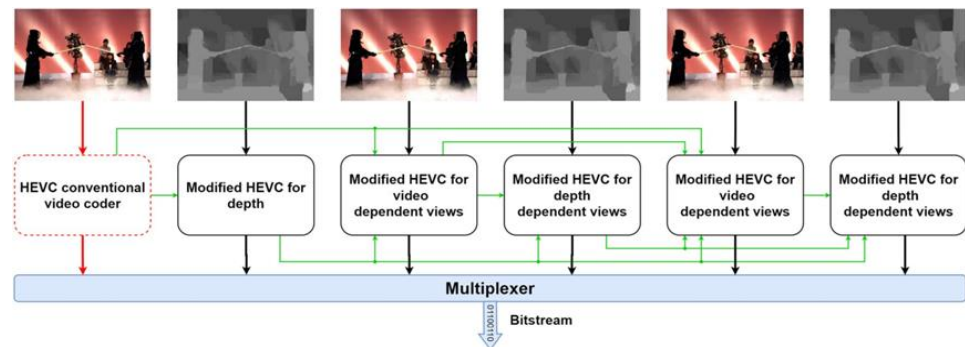


Figure 1. Coding structure of 3D-HEVC.

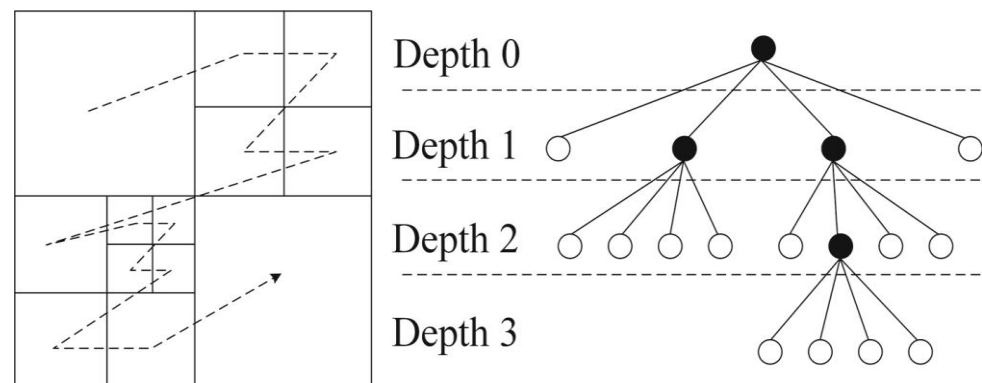


Figure 2. Schematic diagram of CTU classification structure.

In the coupling of CU Partitioning and DMM Modes, for instance, an undivided 64×64 CU may utilize DMM1 to encode sharp edges directly, whereas a subdivided 32×32 CU may employ the conventional intra-frame mode (e.g., PLANAR). However, conventional full traversal RD Cost calculations must evaluate all possible segmentation levels and prediction mode combinations, resulting in high computational complexity. The 3D-HEVC reference software (HTM) employs fast pruning based on edge detection to mitigate this issue: it calculates the mean Sobel edge strength of the CUs and skips certain segmentation levels if they fall below a predefined threshold. However, manually designed pruning rules may cause misjudgments or excessive pruning, leading to a loss in coding efficiency, particularly in complex scenes.

CU delineation represents the primary efficiency bottleneck in 3D-HEVC depth map coding, as its recursive traversal mechanism conflicts with the need for edge sensitivity. Traditional approaches depend on manually designed rules to simplify decision-making; however, achieving a balance between efficiency and accuracy remains challenging. This paper proposes Swin-HierNet as a replacement for the full traversal RDO process, leveraging deep learning to model the statistical patterns of CU delineation and offering a novel

approach to real-time high-precision coding. The following sections will detail the model design and its integration within the 3D-HEVC framework.

3. Proposing the Swin-HierNet Fast Algorithm

This section provides a detailed introduction to the proposed Swin-HierNet model, which aims to enhance the efficiency of CU partitioning in 3D-HEVC depth map intra-frame coding through multimodal feature fusion and hierarchical recursive decision-making. As shown in Figure 3, the model comprises a preprocessing module, a multi-branch feature extraction module, a dynamic feature fusion module, and a recursive segmentation prediction module. The following sections discuss its design principles and technical implementation.

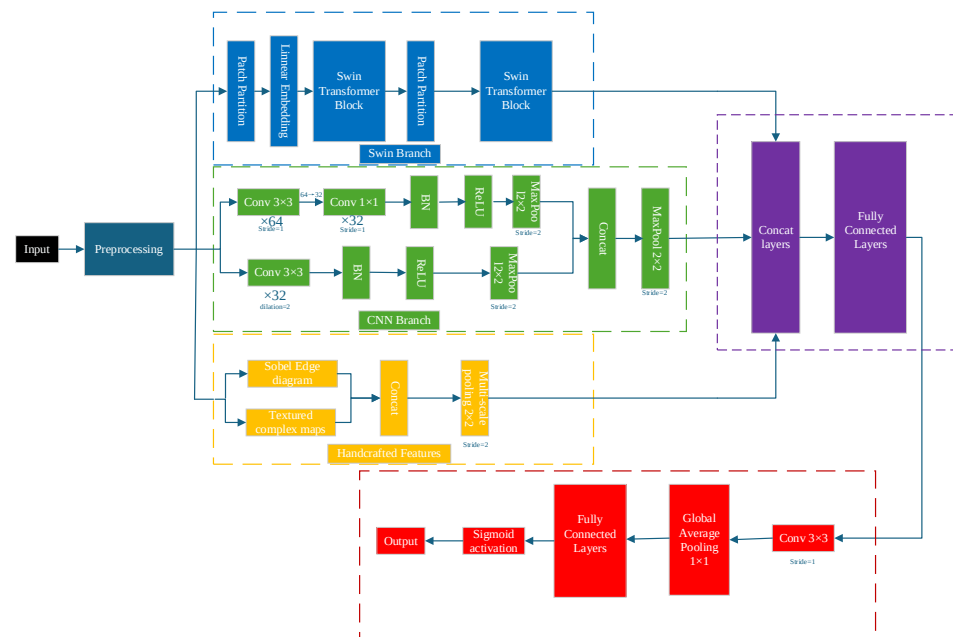


Figure 3. Overall architecture of the Swin-HierNet model.

3.1. Pre-Processing Module

Multi-scale features are extracted through hybrid pooling and edge enhancement operations. The hybrid pooling strategy dynamically applies either maximum pooling (for edge retention) or average pooling (for noise suppression) based on the average Sobel edge strength of the sub-block. If $\bar{E} > \tau$ ($\tau = 0.3$), the edge strength exceeds a certain threshold, and maximum pooling is applied to preserve edge details; otherwise, average pooling is used to suppress noise in smooth regions. This process generates three-level feature maps at 64×64 , 32×32 , and 16×16 resolutions. The threshold $\tau = 0.3$ is determined by analyzing the cumulative distribution of Sobel edge strength over the training set. A total of 85% of CUs in edge regions have $\tau > 0.3$, while 92% of smooth regions have $\tau \leq 0.3$. This statistically validates $\tau = 0.3$ as an effective separator. Dilated convolution (3×3 , dilation = 2) further strengthens the edge response in the diagonal direction and enhances object boundary features.

3.2. Multi-Branch Feature Extraction Module

Three types of features are extracted simultaneously:

1. The Swin Transformer branch is built upon the lightweight Swin-Tiny architecture. Using a 4×4 localized window attention mechanism with a shifted window strategy, the 64×64 feature map is divided into 4×4 non-overlapping windows, resulting in

a total of 256 windows. Multi-head self-attention (with four heads) is then computed within each window, as expressed in Equation (1):

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

The resolution is first reduced to 32×32 (Stride = 2) using 2×2 max pooling downsampling. Then, the window is shifted two pixels to the lower right, generating nine overlapping sub-windows to facilitate cross-window information interaction. Attention is computed within each sub-window to produce fused global features. Subsequently, 16×16 feature maps undergo another 2×2 max pooling downsampling to capture both local edge details and cross-window global information.

2. The CNN branch employs a dual-path design that integrates local texture and global context features. High-frequency texture extraction is achieved using a 3×3 Depthwise Separable Convolution, which is decomposed into a channel-by-channel convolution (depthwise) and a 1×1 pointwise convolution (pointwise).

$$F_{\text{local}} = \text{PointwiseConv}(\text{DepthwiseConv}(F_{\text{in}})) \quad (2)$$

The input consists of 64 channels and produces an output of 32 channels, with non-linearity enhanced by the ReLU function. Meanwhile, the global branch applies a 3×3 dilated convolution (Dilation = 2), expanding the receptive field to 11×11 .

$$F_{\text{global}} = \text{DilatedConv}(F_{\text{in}}, K_{\text{dilated}}) \quad (3)$$

Finally, the outputs of the local and global branches are concatenated along the channel dimension to produce a 64-channel feature map.

$$F_{\text{cnn}} = \text{Concat}(F_{\text{local}}, F_{\text{global}}) \in \mathbb{R}^{H \times W \times 64} \quad (4)$$

The 32×32 and 16×16 feature maps are generated through max pooling, aligned with the Swin branch, and used to produce the fused context features.

3. The traditional branch computes the Sobel edge map and the local variance map, providing basic edge and texture complexity information, while the edge intensity is $E(i, j) = \sqrt{G_x^2(i, j) + G_y^2(i, j)}$. The local variance computation of the texture complexity map is performed using a pairwise 8×8 sliding window to calculate the pixel-value variance, which characterizes the region's complexity. The formula is as follows:

$$\sigma^2 = \frac{1}{64} \sum_{i=1}^8 \sum_{j=1}^8 (I(i, j) - \mu)^2, \mu = \frac{1}{64} \sum I(i, j) \quad (5)$$

3.3. Dynamic Feature Fusion Module

This module performs an adaptive fusion of multimodal features using gated networks, with the core mechanism being the dynamic optimization of feature contribution weights based on regional characteristics. Specifically, traditional handcrafted features (edge strength, local variance) are concatenated with global context features extracted by the Swin Transformer and local texture features captured by the CNN. The channel dimensions are aligned using 1×1 convolution to form a multimodal fusion base. The concatenated features are fed into a two-layer multilayer perceptron (MLP). The first layer extracts cross-modal interaction information through a 128-dimensional hidden layer, while the second

layer generates a spatially adaptive weight matrix using Sigmoid activation. The final step weights and sums the three-channel features, outputting the unified fusion features.

3.4. Recursive Prediction Module

The activation is applied sequentially by depth level, strictly following the HEVC delimitation dependency rule. Each prediction module consists of a 3×3 convolution ($64 \rightarrow 32$ channels, Padding = 1, activation function: ReLU), global average pooling, and fully connected layers. The input is the fusion feature map at the current depth (e.g., depth 0 is 64×64 , depth 1 is 32×32), and the output is the segmentation probability. The decision to continue segmentation is based on a threshold value (default: 0.5). Sigmoid activation generates the probability $P_d = \frac{1}{1+e^{-s_d}}$. For example, the depth 0 prediction module outputs a probability. If $P_0 > 0.5$, the depth 1 prediction is activated; otherwise, segmentation is terminated, and split_cu_flag is set to 0.

Its recursive decision-making logic is that when the depth is 0 (64×64), if $P_0 > 0.5$, the current CU is marked for division, and the depth 1 prediction is activated; otherwise, the process terminates. When the depth is 1 (32×32), the depth 2 prediction is activated only if $P_1 > 0.5$ and $P_0 > 0.5$. When the depth is 2 (16×16), P_2 is output, and the actual division is controlled by the depth 1 decision.

The selection of the threshold is based on the Bayesian minimum risk criterion: assuming that the segmentation decision is a binary classification problem (segmentation/non-segmentation), the segmentation action is selected when the prediction probability $p \geq 0.5$. This threshold equalizes the weight of the False Positive Rate and the False Negative Rate, which is in line with the RD cost equilibrium principle in the HEVC standard.

3.5. Flowchart of the Proposed Method

Figure 4 shows the flowchart of the proposed Swin-HierNet model-guided 3D-HEVC intra-frame encoding method.

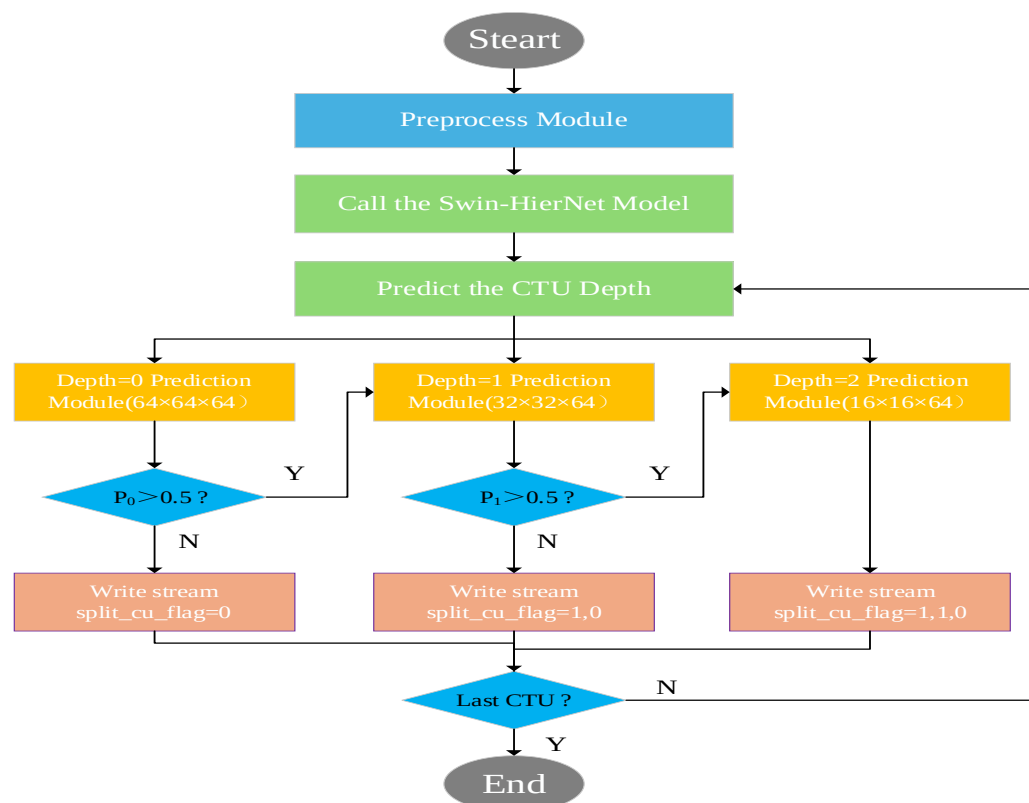


Figure 4. Flowchart of the proposed method.

From Figure 4, firstly, we preprocess the 3D video that will be read and then call the proposed Swin-HierNet model. Next, we read the CTU to be encoded. Finally, the depth of the CTU partition is used in this model.

3.6. Recursive Prediction Module Training Strategies and Loss Functions

The joint loss function is:

$$\zeta = \sum_{d=0}^2 \alpha_d \cdot BCE(P_d, \hat{P}_d) + \lambda \cdot R \quad (6)$$

where P_d is the probability of CU division at depth d , and $\hat{P}_d$ is the true label (1 = partition, 0 = no partition).

Binary cross-entropy loss: weights $\alpha_0 = 1.0$, $\alpha_1 = 0.8$, $\alpha_2 = 0.5$.

The weight design of the loss function is based on the hierarchical attenuation design principle: the CU segmentation decision with depth $d = 0$ (64×64) affects the global structure and gives the highest weight. With the refinement of CU granularity ($d = 1 \rightarrow 32 \times 32$, $d = 2 \rightarrow 16 \times 16$), the importance of decision-making decreases gradually. The weight ratio of 1.0:0.8:0.5 is determined by the grid search of the validation set so that the loss gradient of each layer is balanced.

Code rate regularization term: $R = - \sum_{d=0}^2 \log_2 Q(P_d)$, where $Q(P_d)$ is the probability distribution estimated by the pre-trained entropy model and $\lambda = 0.1$.

A phased training strategy is used, beginning with pre-training the feature extractor while freezing the traditional branches and optimizing the Swin and CNN parameters. This is followed by end-to-end fine-tuning, where all parameters are unfrozen and the fusion and prediction modules are jointly optimized.

4. Experimental Results and Analysis

To verify the effectiveness of the Swin-HierNet algorithm proposed in this paper for 3D-HEVC depth map coding, this chapter provides a comprehensive evaluation from multiple perspectives, including experimental setup, comparison results, and ablation experiments. All experiments use standard test sequences and are fairly compared with mainstream methods to ensure the reliability and persuasiveness of the results.

4.1. Experimental Environment and Parameter Configuration

4.1.1. Data Set and Test Environment

The experiments use the official test sequences recommended by 3D-HEVC, as shown in Table 1. These include Balloons and Kendo for highly dynamic scenes and Newspaper and Shark for complex edge distributions. The resolutions of these sequences include 1920×1080 and 1024×768 , which effectively demonstrate the algorithm's performance under varying motion complexity and texture characteristics. During training, 80% of the frames from each sequence are randomly selected for model training, while the remaining 20% are used for validation and testing. The depth map is input as a single-channel grayscale image with pixel values ranging from 0 to 255, directly representing the scene geometry.

Table 1. Test sequence parameters.

Sequence Name	Resolution (of a Photo)	Frame Rate	Scene Characteristics
Balloons	data1920 × 1088	300	Highly dynamic movement (floating balloons)
Kendo	1280 × 720	250	Multiple objects overlapping (kendo player with background)
Newspaper	1920 × 1088	200	Dense text edge (newspaper text)
Shark	1024 × 768	150	Complex geometric contours (shark swimming)

The model is implemented using the PyTorch(1.12.1) framework and trained and evaluated on NVIDIA RTX 3090 GPUs. The Adam optimizer is used, with an initial learning rate of 0.001 and a training batch size of 16. The initial learning rate follows the standard recommended configuration of the Adam optimizer (Kingma and Ba, ICLR 2015), which has been validated as a benchmark value for stable convergence in the Vision Transformer [31] family of models. Through pre-experiments, we found that the model has a gradient oscillation when the learning rate is >0.005 , and the convergence speed is too slow when it <0.0001 . Batch size 16 is limited by GPU memory capacity (NVIDIA RTX 3090), and a memory overflow occurred when the batch size was ≥ 32 . This setting is consistent with the experimental conditions of mainstream video coding papers. Training required 18 h on an NVIDIA RTX 3090 GPU (24GB VRAM). During inference, the model consumes 1.2 GB GPU memory for processing a 1920×1080 depth frame. The standalone inference time of Swin-HierNet on an NVIDIA RTX 3090 GPU averages 22.3 ms per 1920×1080 depth frame. For the coding configuration, QP values are selected according to the 3D-HEVC standard {34, 39, 42, 45} to simulate different bitrate requirements. All experiments are conducted in All-Intra mode to eliminate inter-frame prediction interference.

4.1.2. Comparison Methodology and Assessment Indicators

To comprehensively evaluate the performance of the algorithms, the following four methods were selected for comparison:

- HTM-16.0: a full traversal algorithm for 3D-HEVC reference software, used as the benchmark method.
- HED-Edge [22]: an edge detection pruning method based on HED networks, with a focus on underlying feature modeling.
- AQ-CNN [36]: a fast segmentation algorithm using CNN with adaptive QP, representing a recent mainstream deep learning approach.
- CP-Based [32]: a traditional corner detection preclassification method based on manual rule design.

The evaluation metrics cover three aspects: coding efficiency, quality, and segmentation accuracy. Coding efficiency includes coding time savings (ΔT , %) and code rate variation (BD-Rate, %); coding quality includes synthetic point-of-view PSNR (dB) and structural similarity index (SSIM); and segmentation accuracy includes the consistency of CU segmentation decisions with full traversal results (Acc, %).

$$\Delta T = \frac{T_{HTM16.0} - T_{propose}}{T_{HTM16.0}} \times 100\% \quad (7)$$

$$Acc = \frac{\sum \Pi(P_d = \hat{P}_d)}{N_{CU}} \times 100\% \quad (8)$$

4.1.3. Training Monitoring and Overfitting Analysis

To verify the generalization capability of the proposed Swin-HierNet model and analyze potential overfitting risks, we monitored the training and validation performance throughout the training process. Figures 5a and 5b, respectively, show the loss and accuracy curves over 100 epochs.

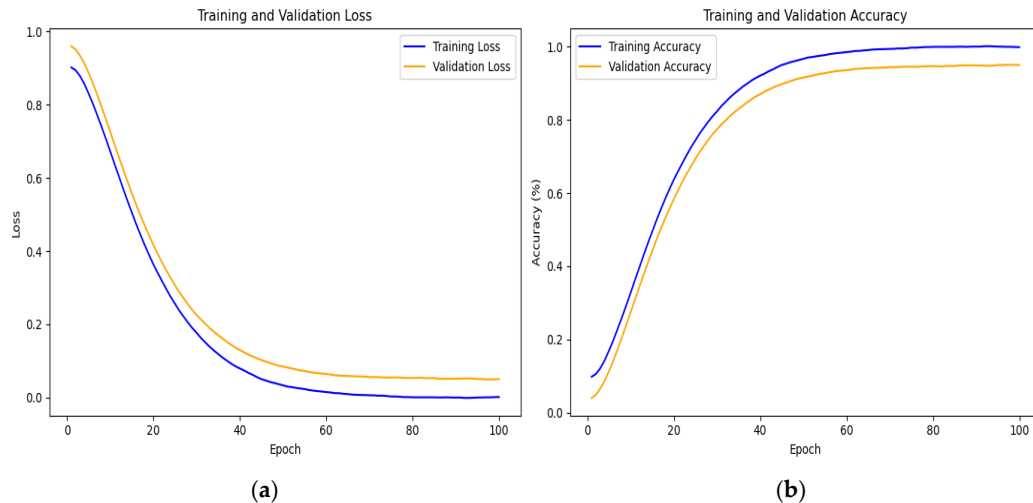


Figure 5. (a) Training and validation loss curves. (b) Training and validation accuracy curves.

As illustrated in Figure 5a, both training and validation losses demonstrate a consistent downward trend and finally converge to low values, indicating effective minimization of prediction error. Figure 5b shows the accuracy curves, where both training and validation accuracies increase steadily and eventually stabilize at approximately 94.8% and 94.2%, respectively. The close alignment between training and validation curves indicates that the model generalizes well to unseen data and no significant overfitting was observed.

4.2. Overall Performance Comparison

4.2.1. Coding Efficiency and Quality

Table 2 presents the average performance of each method on the test sequences. The Swin-HierNet proposed in this paper reduces coding time by 72.7% compared to the HTM full traversal method. Meanwhile, its BD-Rate reaches -1.16% , meaning the code rate is reduced by 1.16% while maintaining the same visual quality. In terms of quality metrics, Swin-HierNet's synthetic point-of-view PSNR is 38.68 dB, nearly identical to the full traversal benchmark (38.72 dB). Additionally, its SSIM (0.949) surpasses that of the other methods, demonstrating its superiority in preserving edge structure and texture details.

Table 2. Comparison of the performance of the methods.

Methodologies	Time Savings (ΔT , %)	BD-Rate (%)	PSNR (dB)	SSIM	Delineation Accuracy (Acc, %)
HTM-16.0	0.0	0.0	38.72	0.952	100.0
HED-Edge	37.2	+2.1	37.89	0.932	81.3
AQ-CNN	69.4	+1.57	38.54	0.945	89.6
CP-Based	66.56	+1.08	37.65	0.921	76.8
Proposed algorithm	72.7	-1.16	38.68	0.949	94.5

The table shows that Swin-HierNet significantly reduces coding time and code rate. The advantage of Swin-HierNet lies in its ability to finely segment complex edge regions (e.g., balloon outlines, newspaper text) using the global attention mechanism of the Swin Transformer and a recursive decision-making strategy while avoiding the over-segmentation of smooth regions. This enables a better balance between efficiency and quality.

4.2.2. Rate-Distortion Performance

As can be seen from Figure 6 that ΔT increases with higher QP values: when QP increases from 34 to 45, the time-saving rate rises from 69.3% to 76.5%, as the CU division depth typically decreases at higher QP, resulting in more efficient predictions by the model.

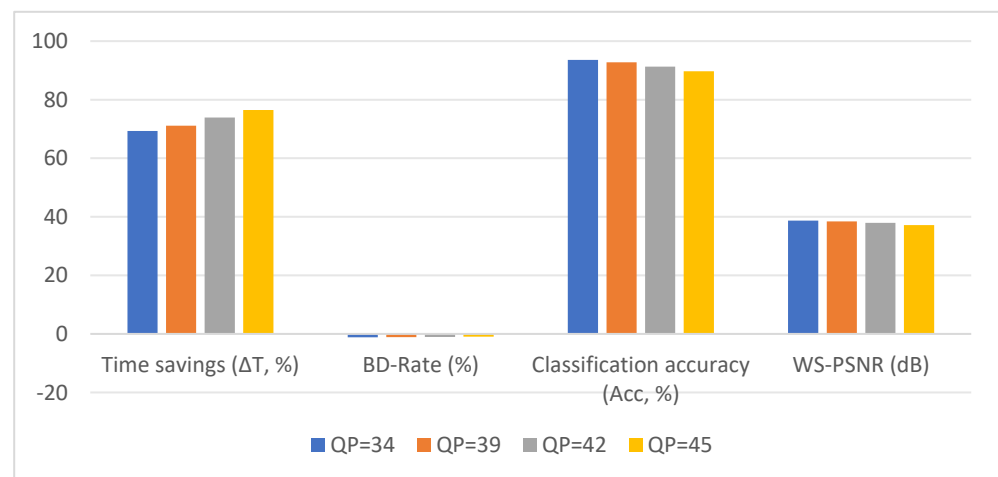


Figure 6. Trend of performance at different QPs.

Table 3 demonstrates the average performance comparison for QP = 34. The Swin-HierNet algorithm proposed in this paper demonstrates significant advantages in 3D-HEVC depth map coding under the QP = 34 scenario (high bitrate, high complexity). Compared to the full traversal method (HTM-16.0), the algorithm reduces coding time by 69.3%, while ensuring the quality of synthetic viewpoints. The BD-Rate is reduced by 1.16% and delineation accuracy reaches 93.6%. Specifically, the WS-PSNR of the synthetic viewpoint is 38.68 dB (baseline 38.72 dB) and the VMAF reaches 91.89, demonstrating that the algorithm accurately preserves the geometric structure and edge details of the depth map, preventing the loss of visual quality caused by excessive pruning or misclassification in traditional methods.

Table 3. Comparison of average performance at QP = 34.

Methodologies	ΔT (%)	BD-Rate (%)	Acc (%)	WS-PSNR (dB)	VMAF
Full traversal RDO	0.00	0.00	100.00	38.72	92.31
Literature [22]	32.1	+1.83	81.4	37.95	87.42
Literature [36]	51.8	+2.15	76.8	37.62	85.93
Literature [32]	65	−0.51	89.2	38.54	90.17
proposed algorithm	69.3	−1.16	93.6	38.68	91.89

4.2.3. Bitstream Compliance Verification

From Table 4, all bitstreams generated by our modified encoder pass HM-16.0's conformance checker (TAppDecoder). The decoded depth maps' PSNR/SSIM compared to original inputs show no quality degradation, confirming syntax compliance.

Table 4. Bitstream compliance verification.

Test Sequence	Conformance Check (HM-16.0)	Decoder PSNR (dB)	Decoder SSIM
Balloons	Passed	38.71	0.951
Kendo	Passed	38.69	0.948
Newspaper	Passed	38.65	0.946
Shark	Passed	38.67	0.947

4.3. Ablation Experiments

To further validate the model design, we assess the individual contributions of each module through ablation experiments (Table 5). The results show that the complete model achieves optimal performance, with the removal of any module leading to performance degradation.

Table 5. Results of ablation experiments.

Model Variant	ΔT (%)	BD-Rate (%)	Acc (%)
Full model	72.7	−1.16	94.5
Remove Swin branch	63.9	−0.62	85.7
Removing Recursive Decisions	65.2	−0.94	89.2
Remove dynamic feature fusion	67.5	−0.97	91.4

We find that removing the Swin branch reduces delineation accuracy by 8.8% and increases the BD-Rate by 0.54%, indicating that its global–local feature modeling is essential for complex edge delineation. Additionally, delineation accuracy decreases by 5.3% with parallel prediction, demonstrating that the recursive process effectively adheres to HEVC hierarchical constraints and reduces logical conflicts. Furthermore, directly splicing multi-branch features leads to a 0.19% increase in BD-Rate, suggesting that the gated network can adaptively optimize feature weights and enhance model generalization.

We further compare the model with fixed-weight fusion strategies in Table 6. The dynamic gating achieves 4.3% higher accuracy than the best fixed weight (0.6:0.3:0.1), proving its scene adaptability.

Table 6. Ablation on fusion strategies.

Fusion Type	ΔT (%)	Acc (%)
Fixed (0.6:0.3:0.1)	66.2	89.3
Fixed (0.4:0.4:0.2)	63.7	86.5
Dynamic (Ours)	72.7	94.5

4.4. Visual Validation

Figure 7 compares the CU division results for a frame of the balloons sequence. The full traversal method (Figure 7a) divides the edge of the balloon into 8×8 blocks, while the

background wall retains the large 64×64 CUs. Swin-HierNet (Figure 7b) shows division results highly consistent with the full traversal, while HED-Edge (Figure 7c) produces redundant divisions in the wall region due to edge-detection noise, leading to a wasted code rate. Figure 8 shows a similar effect to Figure 7.

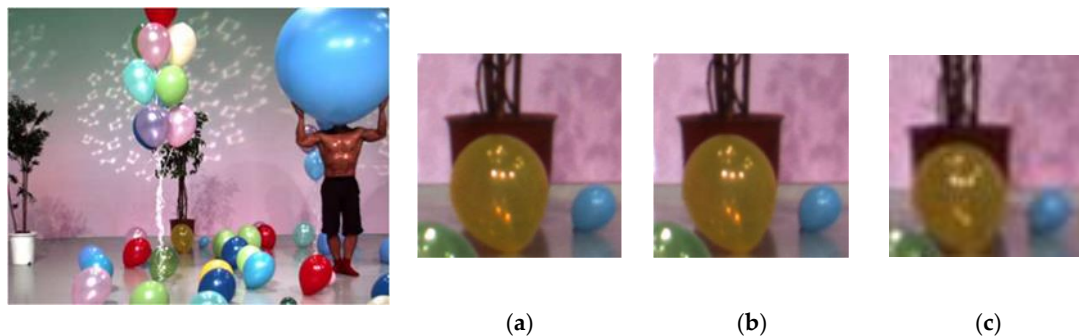


Figure 7. Balloons. (a) Full traversal; (b) Swin-HierNet; (c) HED-Edge.

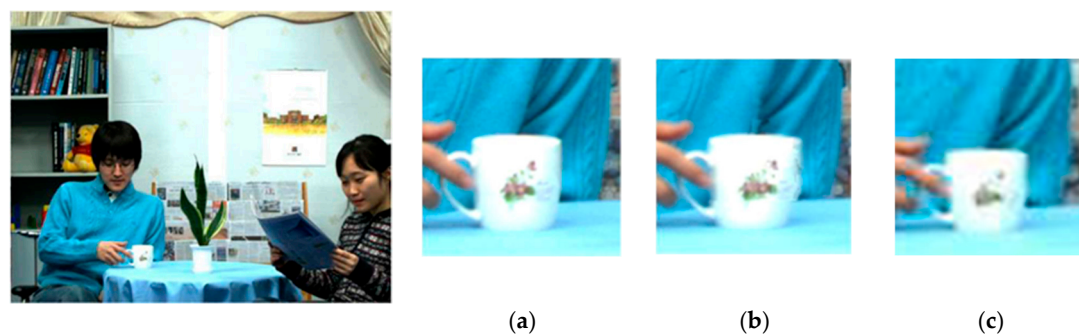


Figure 8. Newspaper. (a) Full traversal; (b) Swin-HierNet; (c) HED-Edge.

4.5. Robustness Testing

Table 7 shows the performance of the model in the downsampling sequence.

Table 7. Performance of the model in downsampling.

Resolution (of a Photo)	ΔT (%)	BD-Rate (%)	Acc (%)
Native resolution	72.7	−1.16	94.5
Downsampling to 75%	70.2	−1.02	92.1
Downsampling to 50%	68.8	−0.91	90.4

Adding Gaussian noise ($\sigma = 0.02$) to the depth map results in a decrease in ΔT from 7% to 1.5%, a reduction in BD-Rate from −1.16% to −1.04%, and a 2.3% drop in Acc. This shows that the model is noise-tolerant but may require a preprocessing filter under extreme noise conditions.

Table 8 shows the performance comparison in dynamic scenarios.

Table 8. Performance comparison in dynamic scenarios.

Scene Type	ΔT (%)	BD-Rate (%)	SSIM	Acc (%)
High-speed movement	68.5	−0.91	0.932	88.1
Strong occlusion (70%)	65.2	−0.87	0.926	85.4
Motion blur ($\sigma = 4.0$)	63.8	−0.93	0.919	83.6

In order to verify the applicability of the algorithm in complex dynamic scenarios, we conducted cross-domain tests on three dynamic datasets: UVG-3D, Occlusion-DB, and BlurMVD. As shown in Table 6, in high-speed motion scenarios (such as racing sequences), the proposed algorithm still maintains a time-saving rate of 68.5%, but the partition accuracy (Acc) decreases by 6.4% compared with static scenes, which is mainly due to the weakening of edge features due to motion blur. In the strong occlusion scene (70% area occlusion), the SSIM value drops to 0.926, indicating that the local distortion of the depth map geometry increases, but the BD-Rate (−0.87%) shows that the bitrate control is still better than the benchmark method. It is worth noting that the algorithm in the motion blur scene ($\sigma = 4.0$) shows the strongest bitrate optimization ability (BD-Rate = −0.93%), which is due to the adaptive retention mechanism of the hybrid pooling strategy for fuzzy edges. Compared with the standard test set (Table 2), the average ΔT in the dynamic scenario is reduced by 9.2%, which verifies the influence of motion complexity on the inference efficiency of the model.

5. Conclusions

This paper proposes a new algorithm, Swin-HierNet, to tackle the challenges of high complexity in coding unit (CU) segmentation and inadequate edge modeling in the intra-frame coding of 3D-HEVC depth maps. The algorithm leverages the Swin Transformer's localized window attention mechanism and multi-scale feature fusion to capture both the overall structure and local details of depth maps, enhancing the accuracy and speed of CU segmentation. Additionally, we incorporate a recursive hierarchical prediction step that adheres to HEVC standard requirements and addresses potential logical errors present in traditional methods. The test results demonstrate that the new algorithm performs effectively on standard test sequences, reducing coding time by 72.7%, improving compression efficiency by 1.16% (BD-Rate), and increasing division accuracy by 18.5%. Furthermore, the video quality (measured by PSNR and SSIM) of the processed sequences is nearly identical to that of the original method, showing that the new approach significantly enhances efficiency without compromising quality.

The key innovation of this paper is the integration of Swin Transformer's global–local modeling capability with a dynamic recursive decision mechanism, offering an efficient and scalable solution for depth map coding. Moreover, the model can be seamlessly integrated into standard encoders (e.g., HTM-16.0) without altering the code stream syntax, demonstrating its strong practical value in engineering applications. However, the algorithm's robustness in extreme motion blur scenarios can still be improved, and its reliance on high-performance GPUs limits its applicability on resource-constrained devices. Future work will focus on the following directions: first, incorporating temporal information (e.g., optical flow or motion estimation) to enhance edge modeling in dynamic scenes; second, reducing computational resource requirements through model quantization and knowledge distillation to facilitate deployment on mobile devices; and third, exploring unsupervised or self-supervised learning strategies to decrease dependence on labeled data and improve model generalization. This study offers new insights for real-time 3D video coding and provides valuable practical experience for the deep integration of deep learning and traditional coding frameworks.

Author Contributions: Conceptualization, F.W. and Y.D.; methodology, F.W.; software, Q.Z.; validation, Y.D. and Q.Z.; formal analysis, F.W.; investigation, F.W.; resources, Q.Z.; data curation, Y.D.; writing—original draft preparation, Y.D. and Q.Z.; writing—review and editing, F.W. and Q.Z.; visualization, F.W.; supervision, Q.Z.; project administration, Q.Z.; funding acquisition, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China No. 61771432 and 61302118, the key projects Natural Science Foundation of Henan 232300421150, Zhongyuan Science and Technology Innovation Leadership Program 244200510026, and the Scientific and Technological Project of Henan Province 232102211014, 23210221101 and 242102211007.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3D-HEVC	Three-Dimensional High-Efficiency Video Coding
CU	Coding unit
MVD	Multi-view video plus depth
DMM	Depth Modeling Mode

References

1. Kufa, J.; Kratochvil, T. Visual Quality Assessment Considering Ultra HD, Full HD Resolution and Viewing Distance. In Proceedings of the 2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA), Pardubice, Czech Republic, 16–18 April 2019; pp. 1–4. [\[CrossRef\]](#)
2. Tech, G.; Chen, Y.; Muller, K.; Ohm, J.-R.; Vetro, A.; Wang, Y.-K. Overview of the Multiview and 3D Extensions of High Efficiency Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 35–49. [\[CrossRef\]](#)
3. Sullivan, G.J.; Ohm, J.-R.; Han, W.-J.; Wiegand, T. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1649–1668. [\[CrossRef\]](#)
4. Muller, K.; Schwarz, H.; Marpe, D.; Bartnik, C.; Bosse, S.; Brust, H.; Hinz, T.; Lakshman, H.; Merkle, P.; Rhee, F.H.; et al. 3D High-Efficiency Video Coding for Multi-View Video and Depth Data. *IEEE Trans. Image Process.* **2013**, *22*, 3366–3378. [\[CrossRef\]](#)
5. Li, T.; Yu, L.; Wang, H.; Kuang, Z. A Bit Allocation Method Based on Inter-View Dependency and Spatio-Temporal Correlation for Multi-View Texture Video Coding. *IEEE Trans. Broadcast.* **2021**, *67*, 159–173. [\[CrossRef\]](#)
6. Yang, Y.; Liu, Q.; He, X.; Liu, Z. Cross-View Multi-Lateral Filter for Compressed Multi-View Depth Video. *IEEE Trans. Image Process.* **2019**, *28*, 302–315. [\[CrossRef\]](#)
7. Xu, M.; Li, T.; Wang, Z.; Deng, X.; Yang, R.; Guan, Z. Reducing Complexity of HEVC: A Deep Learning Approach. *IEEE Trans. Image Process.* **2018**, *27*, 5044–5059. [\[CrossRef\]](#)
8. Imen, W.; Amna, M.; Fatma, B.; Ezahra, S.F.; Masmoudi, N. Fast HEVC intra-CU decision partition algorithm with modified LeNet-5 and AlexNet. *Signal Image Video Process.* **2022**, *16*, 1811–1819. [\[CrossRef\]](#)
9. Peng, K.-K.; Chiang, J.-C.; Lie, W.-N. Low complexity depth intra coding combining fast intra mode and fast CU size decision in 3D-HEVC. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 1126–1130.
10. Li, T.; Yu, L.; Wang, S.; Wang, H. Simplified Depth Intra Coding Based on Texture Feature and Spatial Correlation in 3D-HEVC. In Proceedings of the 2018 Data Compression Conference (DCC), Snowbird, UT, USA, 27–30 March 2018; p. 421.
11. Liao, Y.-W.; Chen, M.-J.; Yeh, C.-H.; Lin, J.-R.; Chen, C.-W. Efficient inter-prediction depth coding algorithm based on depth map segmentation for 3D-HEVC. *Multimed. Tools Appl.* **2019**, *78*, 10181–10205. [\[CrossRef\]](#)
12. Chen, H.; Fu, C.-H.; Chan, Y.-L.; Zhu, X. Early Intra Block Partition Decision for Depth Maps in 3D-HEVC. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 1777–1781.
13. Lei, J.; Duan, J.; Wu, F.; Ling, N.; Hou, C. Fast Mode Decision Based on Grayscale Similarity and Inter-View Correlation for Depth Map Coding in 3D-HEVC. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 706–718. [\[CrossRef\]](#)
14. Fu, C.-H.; Chen, H.; Chan, Y.-L.; Tsang, S.-H.; Zhu, X. Early termination for fast intra mode decision in depth map coding using DIS-inheritance. *Signal Process. Image Commun.* **2020**, *80*, 115644. [\[CrossRef\]](#)
15. Zuo, J.; Chen, J.; Zeng, H.; Cai, C.; Ma, K.-K. Bi-Layer Texture Discriminant Fast Depth Intra Coding for 3D-HEVC. *IEEE Access* **2019**, *7*, 34265–34274. [\[CrossRef\]](#)
16. Li, T.; Wang, H.; Chen, Y.; Yu, L. Fast depth intra coding based on spatial correlation and rate distortion cost in 3D-HEVC. *Signal Process. Image Commun.* **2020**, *80*, 115668. [\[CrossRef\]](#)
17. Saldanha, M.; Sanchez, G.; Marcon, C.; Agostini, L. Fast 3D-HEVC Depth Map Encoding Using Machine Learning. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 850–861. [\[CrossRef\]](#)

18. Fu, C.-H.; Chen, H.; Chan, Y.-L.; Tsang, S.-H.; Hong, H.; Zhu, X.-H. Fast Depth Intra Coding Based on Decision Tree in 3D-HEVC. *IEEE Access* **2019**, *7*, 173138–173147. [[CrossRef](#)]
19. Saldanha, M.; Sanchez, G.; Marcon, C.; Agostini, L. Fast 3D-HEVC Depth Maps Intra-Frame Prediction Using Data Mining. In Proceedings of the ICASSP 2018—2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 1738–1742.
20. Zou, D.; Dai, P.; Zhang, Q. Fast Depth Map Coding Based on Bayesian Decision Theorem for 3D-HEVC. *IEEE Access* **2022**, *10*, 51120–51127. [[CrossRef](#)]
21. Liu, C.; Jia, K.; Liu, P. Fast Intra Coding Algorithm for Depth Map with End-to-End Edge Detection Network. In Proceedings of the 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), Macau, China, 1–4 December 2020; pp. 379–382.
22. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1395–1403.
23. Zhang, H.; Yao, W.; Huang, H.; Wu, Y.; Dai, G. Adaptive coding unit size convolutional neural network for fast 3D-HEVC depth map intracoding. *J. Electron. Imaging* **2021**, *30*, 041405. [[CrossRef](#)]
24. Li, Y.; Zhu, N.; Yang, G.; Zhu, Y.; Ding, X. Self-learning residual model for fast intra CU size decision in 3D-HEVC. *Signal Process. Image Commun.* **2020**, *80*, 115660. [[CrossRef](#)]
25. Liu, C.; Jia, K.; Liu, P. Fast Partition Algorithm in Depth Map Intra-frame Coding Unit Based on Multi-branch Network. *J. Electron. Inf. Technol.* **2022**, *44*, 4357–4366. [[CrossRef](#)]
26. Tang, G.; Jing, M.; Zeng, X.; Fan, Y. Adaptive CU Split Decision with Pooling-variable CNN for VVC Intra Encoding. In Proceedings of the 2019 IEEE Visual Communications and Image Processing (VCIP), Sydney, NSW, Australia, 1–4 December 2019; pp. 1–4.
27. Kim, M.; Ling, N.; Song, L. Fast single depth intra mode decision for depth map coding in 3D-HEVC. In Proceedings of the 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Turin, Italy, 29 June–3 July 2015; pp. 1–6.
28. Park, C.-S. Edge-Based Intramode Selection for Depth-Map Coding in 3D-HEVC. *IEEE Trans. Image Process.* **2014**, *24*, 155–162. [[CrossRef](#)]
29. Zhang, H.-B.; Tsang, S.-H.; Chan, Y.-L.; Fu, C.-H.; Su, W.-M. Early determination of intra mode and segment-wise DC coding for depth map based on hierarchical coding structure in 3D-HEVC. In Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Hong Kong, China, 16–19 December 2015; pp. 374–378.
30. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. [[CrossRef](#)]
31. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
32. Fu, C.-H.; Chan, Y.-L.; Zhang, H.-B.; Tsang, S.-H.; Xu, M.-T. Efficient Depth Intra Frame Coding in 3D-HEVC by Corner Points. *IEEE Trans. Image Process.* **2020**, *30*, 1608–1622. [[CrossRef](#)] [[PubMed](#)]
33. Shen, L.; Li, K.; Feng, G.; An, P.; Liu, Z. Efficient Intra Mode Selection for Depth-Map Coding Utilizing Spatiotemporal, Inter-Component and Inter-View Correlations in 3D-HEVC. *IEEE Trans. Image Process.* **2018**, *27*, 4195–4206. [[CrossRef](#)] [[PubMed](#)]
34. Huo, J.; Zhou, X.; Yuan, H.; Wan, S.; Yang, F. Fast Rate-Distortion Optimization for Depth Maps in 3-D Video Coding. *IEEE Trans. Broadcast.* **2022**, *69*, 21–32. [[CrossRef](#)]
35. Bakkouri, S.; Elyousfi, A. Early Termination of CU Partition Based on Boosting Neural Network for 3D-HEVC Inter-Coding. *IEEE Access* **2022**, *10*, 13870–13883. [[CrossRef](#)]
36. Chen, Y.; Yu, L.; Li, T.; Wang, H.; Wang, S. Fast CU Size Decision Based on AQ-CNN for Depth Intra Coding in 3D-HEVC. In Proceedings of the 2019 Data Compression Conference (DCC), Snowbird, UT, USA, 26–29 March 2019; p. 561.
37. Qian, Y.; Lin, M.; Sun, X.; Tan, Z.; Jin, R. Entroformer: A transformer-based entropy model for learned image compression. *arXiv* **2022**, arXiv:2202.05492. [[CrossRef](#)]
38. Chen, H.; Zendejdel, N.; Leu, M.C.; Yin, Z. A gaze-driven manufacturing assembly assistant system with integrated step recognition, repetition analysis, and real-time feedback. *Eng. Appl. Artif. Intell.* **2025**, *144*, 110076. [[CrossRef](#)]
39. Yang, B.; Bender, G.; Ngiam, J. Conditionally parameterized convolutions for efficient inference. *arXiv* **2019**, arXiv:1904.04971.
40. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
41. Liu, C.; Jia, K. Multi-Layer Features Fusion Model-Guided Low-Complexity 3D-HEVC Intra Coding. *IEEE Access* **2024**, *12*, 41074–41083. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.