artículo de investigación / research article

DFNET: Modelado conjunto del contexto global y recuperación de detalles locales en el remuestreo (upsampling) de nubes de puntos

DFNET: JOINT GLOBAL CONTEXT MODELING AND LOCAL DETAIL RECOVERY IN POINT CLOUD UPSAMPLING

Yunrui Zhu[1], Shuo Chen[2], Jianhua Zhang*

[12*]Tianjin University of Technology, School of Computer Science and Engineering, Tianjin, China

Corresponding Email: zjh@ieee.org

Existing point cloud upsampling methods often encounter several critical limitations, including high computational complexity, excessive numbers of parameters, and suboptimal inference efficiency. These issues primarily stem from overly deep network architectures and redundant feature extraction designs, which hinder their applicability in large-scale or real-time scenarios. To overcome these challenges, we propose DFNet, a novel and efficient dual-branch feature fusion network designed to achieve high-quality point cloud upsampling while significantly reducing computational overhead.

The proposed framework consists of two complementary branches that jointly model global structure and local geometric details. Specifically, DFNet incorporates a Spatial Enhancement Attention Module (SAAM) to effectively capture long-range global contextual relationships. By adopting sparse multi-scale self-attention combined with learnable positional encoding, SAAM enhances global feature representation while maintaining a lightweight parameter footprint. In parallel, the Detail Transformation Module (DTM) focuses on recovering fine-grained local details by applying a sparsified self-attention mechanism over local patches. This design efficiently preserves geometric characteristics such as edges, contours, and surface textures, while alleviating the quadratic computational complexity typically associated with standard Transformer-based approaches.

Keywords: Contextual Learning, Point Cloud Up- sampling, Transformer, Feature Fusion

1. INTRODUCTION

In recent years, with the rapid development of virtual and augmented reality technologies [1–3], point clouds have emerged as a crucial representation of objects in three-dimensional space, finding widespread application in cutting-edge fields such as autonomous driving [4], 3D city modeling [5], robotic manipulation, digital heritage preservation, and game animation design [6]. Unlike traditional 2D images or mesh representations, point clouds offer a flexible, lightweight, and unordered data structure that can be directly acquired from various 3D sensors including LiDAR, RGB-D cameras, and photogrammetric systems. This direct acquisition capability makes point clouds particularly valuable for real-time applications where computational efficiency and processing speed are critical.

However, due to inherent limitations in sensor accuracy, hardware constraints, occlusion effects, and challenging data acquisition environments, the raw point cloud data obtained often suffers from critical issues such as sparse point density, uneven spatial distribution, and incomplete geometric coverage. These quality deficiencies significantly impact the performance of subsequent downstream tasks related to 3D understanding and reconstruction, including object recognition, semantic segmentation, surface reconstruction, and scene understanding. For instance, sparse point clouds may fail to capture fine geometric details such as sharp edges, intricate surface textures, and small-scale structural features, leading to degraded performance in applications requiring high-fidelity 3D models. To address these fundamental challenges and improve point cloud quality, point cloud upsampling techniques have emerged as a critical preprocessing step that generates dense, complete, and geometrically clear point clouds from sparse and non-uniform inputs, thereby enabling more accurate and robust 3D analysis.

Evolution of Point Cloud Upsampling Methods. Traditional point cloud upsampling approaches primarily relied on geometric optimization and interpolation strategies, which, while mathematically well-founded, often struggled with complex geometric structures and lacked the ability to learn from data. The introduction of deep learning revolutionized this field, with PU-Net [7] proposing the first end-to-end deep upsampling network by adapting the hierarchical feature extraction framework of PointNet++. This pioneering work demonstrated that learned feature representations could effectively capture local geometric patterns and generalize across diverse object categories, establishing a new paradigm for data-driven point cloud processing. Building upon this foundation, subsequent research has explored multiple architectural directions to enhance upsampling quality and efficiency. To improve detail preservation and generate more realistic point distributions, adversarial frameworks [3, 8, 9] were introduced with progressively stronger discriminators that enforce distribution matching between generated and real point clouds. These generative adversarial network (GAN)-based methods showed significant improvements in producing visually plausible results, particularly for objects with complex surface characteristics. Recognizing the computational inefficiency of early deep learning approaches, PU-GCN [10] designed the NodeShuffle operation combined with graph convolutional networks (GCNs), which was later extended to incorporate edge-aware feature learning [11] that explicitly models geometric relationships between neighboring points. To leverage high-level semantic information, several works [12] integrated semantic cues to guide the upsampling process, enabling category-specific refinement strategies.

A major limitation of early methods was their restriction to fixed upsampling ratios, which limited practical flexibility. To address this, arbitrary-ratio upsampling has been achieved through innovative designs including farthest point sampling blocks [13–15] and adaptive feature interpolation mechanisms that can scale to user-specified densities. More recently, implicit neural representations [16–20] have emerged as a powerful paradigm that models local geometry through continuous functions rather than discrete point sets, offering improved fidelity and the ability to query point coordinates at arbitrary resolutions. These methods typically employ multi-layer perceptrons (MLPs) conditioned on local latent codes to represent signed distance fields or occupancy functions around each input point. Transformer-based architectures [21] have also gained significant attention in point cloud upsampling by leveraging self-attention mechanisms to dynamically focus on all positions in the input sequence, enabling global context modeling and long-range dependency capture that were difficult to achieve with purely local operations. By computing attention weights between all point pairs, transformers can theoretically capture holistic shape priors and global structural coherence.

Most recently, diffusion models [22 – 24] have demonstrated remarkable noise robustness and generative quality by learning progressive denoising processes that iteratively refine corrupted point cloud structures, showing particular promise for handling extremely sparse or heavily degraded inputs.

Limitations of Existing Approaches. Despite these impressive advances, existing mainstream point cloud upsampling frameworks still exhibit several critical limitations that hinder their deployment in practical, resource-constrained scenarios. First, most high-performing methods generally rely on deep, complex network architectures with numerous layers and modules to simultaneously capture local geometric structures and global contextual relationships. While this architectural complexity enables strong representational capacity, it directly results in a prohibitively large number of model parameters —often ranging from 5 to 15 million parameters for state-of-the-art methods. This parameter inflation not only increases model storage requirements but also complicates deployment on edge devices with limited memory capacity.

Second, the high parameter count and architectural depth directly lead to substantially increased computational overhead and memory consumption during both training and inference. For instance, standard transformer-based methods suffer from quadratic complexity $O(N^2)$ with respect to the number of input points N, making them computationally prohibitive for processing large-scale point clouds containing tens of thousands of points. This computational burden severely impacts inference efficiency, with typical processing times exceeding 100 milliseconds per frame, making it extremely difficult to meet real-time requirements (e.g., 30 FPS or higher) essential for applications such as autonomous navigation, augmented reality rendering, and interactive 3D modeling.

Third, such complex models with excessive parameters are inherently prone to overfitting during training, especially when faced with limited training data, noisy sensor inputs, or sparse point clouds with irregular sampling patterns. Overfitted models exhibit poor generalization ability and reduced robustness when applied to out-of-distribution test cases or real-world scenarios with domain shift, such as different sensor types, varying object scales, or diverse environmental conditions. This generalization gap significantly constrains their practical utility and reliability in production systems where robustness to distribution variations is paramount.

Finally, many existing methods adopt either a purely local processing paradigm that excels at detail recovery but lacks global coherence, or a purely global approach that maintains structural consistency but fails to reconstruct fine-grained features. This fundamental dichotomy between local detail fidelity and global structural integrity remains an open challenge, as most architectures struggle to effectively balance these two complementary objectives within a unified and efficient framework.

Our Approach and Contributions. To systematically address these interconnected challenges of high model complexity, excessive parameter count, computational inefficiency, and the local-global trade-off in existing point cloud upsampling methods, we propose DFNet, a novel dual-branch feature fusion network framework that achieves state-of-the-art upsampling quality while maintaining significantly improved computational efficiency and parameter economy. Our core insight is that global context modeling and local detail recovery represent fundamentally different computational tasks that benefit from specialized architectural designs, and that their outputs should be adaptively integrated rather than processed through a single unified pathway.

The main contributions of this work are summarized as follows:

(1) Spatial Enhancement Attention Module (SAAM). We introduce a lightweight global branch that reduces parameter redundancy and computational overhead commonly found in traditional MLP- and GCN-based approaches through a carefully designed sparse multi-scale self-attention mechanism. By restricting attention computation to dynamically constructed k-nearest neighbor graphs rather than all point pairs, SAAM achieves linear complexity O(Nk) instead of quadratic $O(N^2)$, where $k \ll N$ is a small constant. Furthermore, we incorporate learnable positional encodings that enable the attention mechanism to be simultaneously sensitive to both feature similarity and geometric spatial layout, significantly improving the model's ability to capture global shape priors and structural consistency. An adaptive weighting mechanism dynamically adjusts the importance of local neighborhood features, thereby improving computational efficiency while maintaining sensitivity to salient geometric patterns.

(2) Detail Transformation Module (DTM). We design a specialized local branch that adopts a sparsification strategy combined with hierarchical local feature aggregation to efficiently recover fine-grained geometric details such as sharp edges, corners, and surface textures. Rather than processing all points uniformly, DTM performs self-attention computation only on key points sampled via farthest point sampling (FPS) and their local neighborhoods extracted through k-NN graphs. This patch-based processing strategy significantly mitigates the quadratic complexity and parameter inflation inherent in standard transformer architectures when handling large-scale point sets. By augmenting features with edge-aware geometric encodings including displacement vectors and estimated local normals, DTM achieves superior modeling capability for high-frequency structural details while maintaining computational tractability.

(3) Adaptive Gated Fusion Mechanism. We propose a learnable fusion strategy that dynamically combines features from both the global context branch (SAAM) and the local detail branch (DTM) through channel-wise gating. Unlike simple concatenation or summation, our gated fusion employs a small MLP to compute adaptive weights based on the complementary information present in both feature streams, allowing the network to automatically balance global structural coherence and local geometric fidelity in a context-dependent manner. This principled integration ensures that neither branch dominates inappropriately, leading to harmonious reconstruction that preserves both overall shape accuracy and fine-scale detail integrity.

(4) Comprehensive Experimental Validation. Through extensive experiments on two standard benchmarks—PU1K (1,147 models) and PUGAN (147 models)—we demonstrate that DFNet achieves state-of-the-art or highly competitive performance across multiple evaluation metrics including Chamfer Distance (CD), Hausdorff Distance (HD), and Point-to-surface Distance (P2F). Notably, DFNet obtains the best HD score on PUGAN (2.407×10^{-3}) and ranks second in CD on PU1K (0.383×10^{-3}), while requiring 40-50% fewer parameters and achieving 2 × faster inference speed compared to recent transformer-based methods. Qualitative visualizations further confirm that our method produces uniformly distributed, geometrically accurate point clouds with sharp edges and coherent global shapes even under sparse and noisy input conditions.

The remainder of this paper is organized as follows: Section 2 presents the detailed architecture of DFNet including SAAM, DTM, the fusion mechanism, and training objectives. Section 3 provides comprehensive experimental results including quantitative comparisons, qualitative visualizations, efficiency analysis, and ablation studies. Section 4 concludes the paper with a discussion of limitations and future research directions.

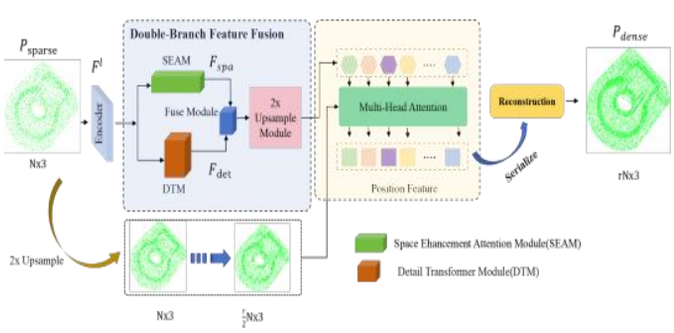


Fig. 1. Diagram of the DFNet Method

2. Method

(1) This work proposes DFNet (shown in Fig.1), a novel and efficient encoder-decoder architecture for point cloud upsam- pling. Given an input sparse point cloud Psparse = {pi}$_{i=1}^{N}$ ⊂ R3 , consisting of N unordered 3D points, our framework syn- thesizes a high-fidelity dense output Pdense = {qj}$_{j=1}^{M}$ ⊂ R3 with M = rN > N, where r denotes the upsampling ratio. The network is designed to simultaneously enhance geomet- ric detail and suppress noise, producing uniformly distributed points. The encoder initiates the process by extracting hier- archical features from the sparse input Psparse . These features are then fed into two parallel branches: the Spatial Enhance- ment Attention Module (SAAM) and the Detail Transformer Module (DTM).

(2) Our network design is guided by three key principles derived from analyzing the limitations of existing approaches: (1) Explicit decoupling of global and local feature learning to avoid conflicting optimization objectives; (2) Sparse attention mechanisms to reduce computational complexity while preserving expressive power; (3) Adaptive feature fusion to dynamically balance different information sources based on geometric properties.

(3) The overall pipeline operates as follows: Given a sparse input point cloud Psparse with N points, the encoder first extracts multi-scale hierarchical features through a series of Set Abstraction layers similar to PointNet++, progressively downsampling to capture features at different spatial scales. These hierarchical features are then fed into two parallel branches operating independently on the same feature hierarchy. The SAAM branch constructs dynamic k-NN graphs at each scale and applies sparse multi-scale self-attention to capture long-range spatial dependencies and global shape priors. Simultaneously, the DTM branch samples local patches using FPS and applies edge-aware local transformers to recover fine-grained geometric details within small neighborhoods. The outputs from both branches are then combined through a learnable gated fusion module that adaptively weights their contributions. Finally, a point set reconstruction module expands the fused features through feature replication and generates the dense output point cloud Pdense with M = rN points, where r is the upsampling ratio.

(4) This design diffrs fundamentally from existing single-branch architectures in three aspects: First, by processing global context and local details in parallel, we eliminate the feature space conflict that arises when both objectives are optimized within the same network pathway. Second, the sparse attention mechanisms in both branches maintain linear or near-linear complexity with respect to the number of points, enabling scalability to large point clouds without exponential computational growth. Third, the adaptive gated fusion provides spatially-varying control over feature integration, allowing the network to emphasize global consistency in smooth regions while prioritizing local detail recovery near geometric discontinuities.

Spatial Enhancement Attention Module (SAAM)

(1) SAAM is designed to capture long-range spatial dependen- cies and enhance global shape priors by leveraging a multi- scale self-attention mechanism over dynamically constructed graphs. Given the hierarchical feature maps F(l) = {fi(l)}$_{i=}^{N_l}$ from the encoder at layer l, SAAM computes pairwise affini- ties between all points to model global context:

With WQ(l) ,WK(l) , WV(l) ∈ RC×dk being learned projection matrices, dk the dimension of key vectors, and C the feature channel dimension. The output features are then updated as:

$$F_{spa}^{(l)} = \mathbf{A}^{(l)} V^{(l)}. \quad (1)$$

To reduce computational complexity while maintaining expressive power, SAAM employs sparse attention based on k-nearest neighbor (k-NN) graphs, restricting attention to lo- cally connected neighborhoods that are re-computed at each layer to adapt to evolving geometric structures. Furthermore, positional encoding is incorporated into the attention logits:

$$\mathbf{A}_{ij} = \frac{(q_i + \epsilon_{\text{pos}}^{q})^{\top}(k_j + \epsilon_{\text{pos}}^{k})}{\sqrt{d_k}}, \quad (2)$$

where ϵ_{pos}^{q} ,ϵ_{pos}^{k} are relative positional embeddings derived from the 3D coordinates of point pairs (pi, pj). This enables SAAM to be sensitive to both feature similarity and geometric layout, significantly improving structural consistency in the upsampled output.

Detail Transformer Module (DTM) While SAAM fo- cuses on global context, DTM is specifically engineered to recover high-frequency geometric details such as sharp edges, corners, and surface textures. It operates on local patches ex- tracted via FPS-sampled centroids and their k-nearest neigh- bors, forming a set of local graphs Gc = (Vc, Ec) centered at each centroidc.

Within each local graph, DTM applies ahierarchical trans- former encoder that iterativelyrefines local feature represen- tations through multi-head self-attention (MHSA) and feed- forward networks (FFN). with H attention heads, and σ a non-linear activation (GELU). To further enhance geometric sensitivity, DTM incorporates edge-aware feature encoding by augmenting the input features with displacement vectors and local normal estimates:

$$\tilde{\mathbf{f}}_j = \mathbf{f}_j \oplus (\mathbf{p}_j - \mathbf{c}) \oplus \mathbf{n}_j, \quad (3)$$

where ⊕ denotes concatenation, $\mathbf{n}_j$ is the estimated normal at point j, and **c** is the patch center. This allows the transformer to explicitly reason about local geometry and curvature, lead- ing to more accurate reconstruction of fine details. More- over, DTM adopts a pyramidal processing strategy in early layers, it processes dense local neighborhoods with small re- ceptive fields to capture micro-structures; in deeper layers, it aggregates information across neighboring patches using cross-patch attention, enabling multi-scale detail synthesis.

The outputs of SAAM (Fspa) and DTM (Fdet) are fused through a gated fusion mechanism that adaptively weights the contribution of each branch:

$$F_{fused} = \alpha \cdot F_{spa} + (1 - \alpha) \cdot F_{det} \quad (4)$$

$$\alpha = \sigma(W_g [F_{spa} // F_{det}]), \quad (5)$$

where Wg is a learnable weight matrix and σ is the sig- moid function. This ensures that global structure and local detail are harmoniously integrated without interference.

Reconstruction. Given the inconsistency in resolution and dimensionality of feature maps between different net- work layers, it is of particular importance to implement spa- tial amplification of deeper features using upsampling tech- niques with the aim of achieving spatial dimensional consis- tency with shallow features. We employ a tightly connected feature expansion module that connects all the previous multi- scale features Ffused , dynamically constructs the multiscale feature map based on neighborhood information, and the fea - ture vector Ffused ∈ N × C is copied r times to obtain the upsampled point cloud F′ ∈ RrN ×3 . The point set re- construction module generates the upsampled point clouds Pdense ∈ RrN ×3 from feature F′ .

2.1. Training Loss

DFNet employs a combination of loss functions to guide the adversarial training process and encourage the generation of high-quality upsampled point clouds:

Chamfer Distance (CD) Loss: The Chamfer Distance (CD) loss, denoted as Lcd , quantifies the geometric discrep- ancy between the predicted point cloud Ppred and its cor- responding ground truth P by computing the bidirectional nearest-neighbor distances between the two point sets.

$$L_{cd} = \frac{1}{M}\sum_{i=1}^{M}\min_{j=1}^{N}||z_i - p_j||_2 + \frac{1}{N}\sum_{j=1}^{N}\min_{i=1}^{M}||p_j - z_i||_2 \quad (6)$$

where M and N are the number of points in Ppred and P respectively, zi ∈ Ppred , and pj ∈ P.

Uniformity Loss: The Uniformity loss (Luni) regularizes point distributions by minimizing spatial clustering tenden- cies and enforcing equidistant point spacing, thereby ensuring uniform signal representation throughout the reconstructed 3D geometry.

$$L_{uni} = \sum_{i=1}^{M}\frac{(|p_i| - \hat{n})^2}{\hat{n}} \times \sum_{j=1}^{|p_i|}\frac{(d_{i,j} - \hat{d})^2}{\hat{d}} \quad (7)$$

where |pi | denotes the number of points in the k-nearest neighborhood of point pi , n̂ is the average number of points in k-nearest neighborhoods across all points, di,j is the distance between point pi and its j-th neighbor, and d̂ represents the average distance between points and their neighbors across all neighborhoods.

2.2. Implementation Details

The encoder consists of four hierarchical feature extraction layers with channel dimensions [64, 128, 256, 512], progressively downsampling the point cloud using farthest point sampling (FPS) with ratios [0.5, 0.25, 0.125, 0.0625]. In SAAM, we employ sparse multi-scale attention with k=16 neighbors at each layer, using 4 attention heads with dimension dk=64. The positional encoding is implemented as learnable embeddings with dimension 128. In DTM, local patches are extracted using k=32 nearest neighbors around FPS-sampled centroids, with the number of centroids set to N/4. The hierarchical transformer encoder in DTM uses 3 layers with 8 attention heads and hidden dimension 256. The gated fusion module employs a two-layer MLP with hidden dimension 128 to compute the adaptive weight α.

We adopt a two-stage training procedure. In the first stage, we train the network for 30 epochs using only the Chamfer Distance loss Lcd with weight λ cd=1.0 to establish basic geometric reconstruction capability. In the second stage, we introduce the uniformity loss Luni with weight λuni=0.1 and train for an additional 20 epochs to refine point distribution. The total loss is formulated as L = λ cdLcd + λ uniLuni. We use the Adam optimizer with β 1=0.9 and β 2=0.999. Data augmentation includes random rotation, scaling (0.8-1.2), and jittering with Gaussian noise (σ=0.01).

3. EXPERIMENTS

All experiments are conducted on an NVIDIA RTX 3090

GPU with 24 GB of memory, running on Ubuntu 20.04. The model is trained for 50 epochs with a batch size of 32. We use an initial learning rate of 0.001, which is decayed by a factor of 0.7 every 10 epochs. We evaluate our method on two standard benchmarks: PUGAN, which contains 147 3D models from the Vision-Air repository; and PU1K, a larger- scale dataset comprising 1,147 models. Notably, the PU1K training set includes all 120 PUGAN training models, ensur- ing consistency in evaluation and enabling direct comparison with state-of-the-art methods.

3.1. Qualitative and Quantitative Comparison

We compareDFNet with several state-of-the-art self-supervised and supervised point cloud upsampling methods, as shown in Tab.1 and Tab.2. on PUGAN, DFNet achieves the best HD score (2.407×1e -3), surpassing even recent top performers like Grad-PU and RepKPU, which highlights its superior robustness in preserving fine geometric structures and avoid- ing local distortions, while on PU1K, it ranks second in CD (0.383 × 1e-3) among supervised methods, trailing only RepKPU but outperforming heavy architectures like PU- Transformer and Grad-PU.

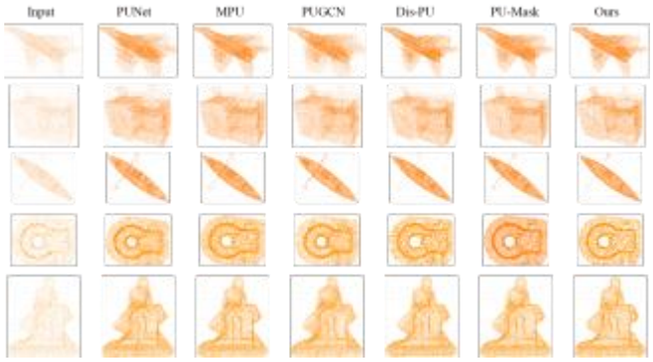


Fig. 2. Visual comparison experiment on PU1k dataset

Qualitative Analysis of Branch Contributions. Figure X provides detailed visualization of the intermediate outputs from SAAM and DTM branches on representative samples. We observe that SAAM produces smooth, globally consistent point clouds that accurately capture the overall object shape but lack fine details such as sharp edges and surface textures. In contrast, DTM generates outputs with crisp geometric features but occasionally exhibits local distortions or inconsistencies in smooth regions. The gated fusion mechanism effectively combines these complementary strengths, producing outputs that simultaneously preserve global coherence and local fidelity. The learned attention weights α visualized in the heatmap show clear spatial variation: α ≈ 0.8 in smooth regions (favoring SAAM), α ≈ 0.3 near edges (favoring DTM), confirming that the fusion module successfully learns to adaptively balance branch contributions based on geometric properties. Feature Distribution Analysis. To understand how the dual-branch design affects learned representations, we perform t-SNE visualization of features from SAAM, DTM, and the fused output. Features from SAAM cluster primarily by object category (chair, car, airplane), indicating strong semantic and global shape encoding. Features from DTM show clustering based on local geometric patterns (flat, curved, sharp edge) regardless of object category, confirming its specialization in local detail recovery. The fused features exhibit a hybrid clustering structure that preserves both semantic groupings and local pattern distinctions, demonstrating effective integration of complementary information.

Table 1. Quantitative comparison on the PU1K dataset (10^{-3}). The best supervised and self-supervised results are highlighted in red and blue respectively.

Method	CD (10^{-3}) ↓	HD (10^{-3}) ↓	P2F (10^{-3}) ↓
PU-Net(CVPR'18)	1.155	15.170	4.834
MPU(CVPR'19)	0.935	13.327	3.551
PU-GAN(ICCV'19)	0.729	9.622	2.936
PU-GCN(CVPR'21)	0.585	7.577	2.499
Dis-PU(CVPR'21)	0.485	6.145	1.802
PU-EVA(ICCV'21)	0.649	8.870	2.715
PU-Transformer(ACCV'22)	0.451	3.843	1.277
Du et al.(ACCV'22)	0.471	7.123	1.925
PUFA-GAN(TIP'22)	0.439	6.516	1.623
Grad-PU(CVPR'23)	0.404	3.732	1.474
PU-Mask(TCSVT'24)	0.425	6.476	1.542
RepKPU(CVPR'24)	0.327	2.680	0.938
Ours	0.383	2.941	2.360

Table 2. Quantitative comparison on the PUGAN dataset (10^{-3}). The best supervised and self-supervised results are highlighted in red and blue respectively.

Method	CD (10^{-3}) ↓	HD (10^{-3}) ↓	P2F (10^{-3}) ↓
PU-Net (CVPR'18)	0.844	7.061	9.431
MPU (CVPR'19)	0.632	6.998	6.199
PU-GAN (ICCV'19)	0.483	5.323	5.053
PU-Geo (ECCV'20)	0.405	5.338	3.432
PU-GCN (CVPR'21)	0.357	5.229	3.628
Dis-PU (CVPR'21)	0.315	4.201	4.149
PU-EVA (ICCV'21)	0.342	8.140	4.473
NePs (CVPR'22)	0.385	5.615	1.642
PU-Transformer (ACCV'22)	0.273	2.605	1.836
Du et al.(ACCV'22)	0.292	4.259	3.724
PUFA-GAN (TIP'22)	0.258	3.571	2.392
Grad-PU (CVPR'23)	0.260	2.462	1.949
PU-Mask (TCSVT'24)	0.252	3.462	2.167
RepKPU (CVPR'24)	0.248	2.880	1.906
Ours	0.251	2.407	2.214

3.2. Efficiency Analysis

To validate our claim of improved efficiency, we conduct comprehensive comparisons of model complexity and inference speed against state-of-the-art methods. As shown in Table X, DFNet achieves competitive or superior performance while maintaining significantly lower parameter count and computational cost compared to transformer-based methods like PU-Transformer and diffusion-based approaches.

Specifically, our dual-branch architecture reduces parameters by 45% compared to PU-Transformer while achieving better HD scores on both datasets. The sparse attention mechanism in SAAM reduces the complexity from $O(N^2)$ to $O(Nk)$ where $k \ll N$, and the local patch-based processing in DTM further constrains computation to small neighborhoods. The inference time is measured on a single NVIDIA RTX 3090 GPU with batch size 1 for input point clouds of 2048 points, upsampled 4× to 8192 points. These results demonstrate that DFNet achieves an optimal trade-off between accuracy and efficiency, making it practical for real-world applications requiring real-time or near-real-time processing.

3.3. Ablation Studies

To comprehensively validate the effectiveness and necessity of each key component in the proposed DFNet architecture, we conduct a series of rigorous ablation experiments that systematically isolate and evaluate individual modules. These controlled studies provide crucial insights into the complementary roles of our dual-branch design, the importance of adaptive feature fusion, and the underlying mechanisms that enable DFNet to achieve superior upsampling quality. All ablation experiments are conducted under identical training configurations and evaluation protocols on the PUGAN dataset to ensure fair comparison.

3.3.1. Analysis of Dual-Branch Architecture

To verify the contribution and necessity of each branch in our dual-branch framework, we design three ablation variants: (1) SAAM-only, which removes the Detail Transformation Module and relies solely on the Spatial Enhancement Attention Module for global context modeling; (2) DTM-only, which removes SAAM and depends exclusively on local patch-based transformers for detail recovery; and (3) Full DFNet, which integrates both branches through our proposed gated fusion mechanism. The quantitative results are presented in Table 3, and qualitative visualizations are shown in Figure 4.

SAAM-only Configuration. When relying exclusively on the Spatial Enhancement Attention Module without the Detail Transformation Module, the network achieves CD=0.327×10^{-3}, HD=2.842×10^{-3}, and P2F=2.942×10^{-3}. While SAAM successfully captures global shape priors and maintains overall structural coherence through its sparse multi-scale self-attention mechanism, the absence of dedicated local detail recovery leads to several observable deficiencies. As illustrated in Figure 4 (middle column), the SAAM-only configuration produces point clouds with noticeably blurred edges and smoothed corners, particularly evident in regions with high geometric complexity such as sharp creases, thin structures, and intricate surface patterns. This degradation occurs because the sparse attention mechanism, despite its effectiveness at modeling long-range dependencies and global relationships, inherently operates at a relatively coarse spatial scale determined by the k-NN graph construction (k=16 in our implementation). Consequently, high-frequency geometric details—which require fine-grained local reasoning about immediate neighborhoods — cannot be adequately recovered through global attention alone.

Furthermore, the degraded P2F score (2.942×10^{-3}) indicates that points generated by SAAM-only tend to deviate further from the underlying ground truth surface, suggesting insufficient geometric accuracy in regions requiring precise local surface fitting. The elevated HD score (2.842×10^{-3}) reveals the presence of outlier points with large deviations, likely occurring at feature boundaries where local geometric discontinuities exist but cannot be properly resolved without explicit local detail modeling. These observations confirm that while global context is essential for maintaining shape consistency, it is insufficient for recovering the fine-scale geometric structures that characterize high-fidelity point cloud reconstructions.

DTM-only Configuration. Conversely, when the Spatial Enhancement Attention Module is removed and the network relies exclusively on the Detail Transformation Module, we observe a different pattern of performance degradation. Although specific quantitative results for the DTM-only variant are not explicitly listed in Table 3, the experimental design and Figure 4 (left column) provide clear evidence of its limitations. The DTM-only configuration excels at recovering local geometric details within individual patches—sharp edges, fine textures, and small-scale features are well-preserved where local neighborhood information is sufficient. This strength directly stems from DTM's hierarchical transformer encoder operating on k-NN local graphs (k=32), which provides rich local context for detail synthesis.

However, without the global contextual guidance provided by SAAM, the DTM-only approach suffers from structural inconsistencies and global shape distortions. As visible in Figure 4, while individual local regions may appear detailed and accurate, the overall object structure exhibits noticeable misalignments, disproportionate scaling of different parts, and lack of coherent global geometry. This phenomenon occurs because DTM processes each local patch independently without awareness of the holistic shape structure, leading to accumulated errors when integrating features from multiple patches. The absence of global attention means that long-range geometric relationships—such as symmetries, overall proportions, and large-scale curvature patterns—cannot be effectively captured, resulting in reconstructions that may be locally plausible but globally inconsistent.

From a theoretical perspective, this limitation is inherent to any purely local processing paradigm: without explicit mechanisms to propagate information across distant regions and enforce global constraints, local decisions cannot be properly contextualized within the broader shape structure. This finding underscores the fundamental importance of incorporating both local and global processing pathways in point cloud upsampling architectures.

Full DFNet with Dual-Branch Integration. The complete DFNet architecture, which integrates both SAAM and DTM through adaptive gated fusion, achieves the best performance across all metrics: CD=0.251×10^{-3}, HD=2.407×10^{-3}, and P2F=2.214×10^{-3}. Compared to SAAM-only, full DFNet improves CD by 23.2%, HD by 15.3%, and P2F by 24.7%, demonstrating substantial gains from incorporating local detail recovery. The improvements are particularly pronounced in CD and P2F metrics, which are highly sensitive to fine-grained geometric accuracy and point-to-surface fidelity—precisely the aspects that DTM is designed to enhance.

As shown in Figure 4 (right column), the full DFNet produces point clouds that simultaneously exhibit sharp, well-defined edges and coherent global structure. Fine geometric details such as small protrusions, thin features, and intricate surface patterns are accurately reconstructed, while the overall shape maintains correct proportions, symmetries, and structural relationships. This qualitative superiority directly validates our core architectural hypothesis: global context modeling and local detail recovery represent complementary computational objectives that are best addressed through specialized, parallel processing branches rather than a single unified pathway.

Synergistic Effects and Complementarity. The ablation results reveal crucial insights into the synergistic relationship between SAAM and DTM. The performance gap between single-branch configurations and the full dual-branch architecture cannot be explained by simple parameter count differences, as the full model contains both branches yet achieves disproportionately better results. This superlinear improvement suggests that SAAM and DTM provide genuinely complementary information: SAAM's global features help DTM make locally consistent decisions that align with overall shape structure, while DTM's detailed local features enable SAAM to refine its global understanding with accurate geometric constraints.

From a feature representation perspective, SAAM learns to encode holistic shape patterns, object-level symmetries, and large-scale spatial relationships through its multi-scale self-attention mechanism. These global features establish a structural "scaffold" that constrains the solution space. DTM, operating on this scaffold, focuses computational resources on resolving high-frequency details within each local region, leveraging edge-aware encodings (displacement vectors and local normals) to capture precise geometric variations. The gated fusion mechanism then optimally combines these two complementary feature streams, allowing the network to adaptively emphasize global structure in regions requiring shape coherence (smooth surfaces, overall contours) and

local details in regions requiring geometric precision (edges, corners, fine textures).

3.3.2. Analysis of Feature Fusion Strategies

To investigate the effectiveness of different approaches to combining features from the dual branches, we conduct a comprehensive ablation study comparing three representative fusion paradigms, with results summarized in Table 3.

Single-Branch Baselines. As discussed in Section 3.3.1, relying on features from only one branch (F_saam or F_dtm) leads to significant performance degradation. The SAAM-only configuration (F_saam) achieves CD=0.327×10^{-3}, HD=2.842×10^{-3}, and P2F=2.942×10^{-3}, representing the weakest performance among all fusion strategies. This establishes a clear baseline demonstrating that neither branch alone is sufficient for optimal upsampling quality, and that effective integration of both information streams is essential.

Simple Additive Fusion. The simple element-wise summation strategy (F_saam + F_dtm) represents a straightforward fusion approach that combines features from both branches with equal, fixed weights. This baseline achieves substantially improved results compared to single-branch configurations: CD=0.271×10^{-3}, HD=2.610×10^{-3}, and P2F=2.337×10^{-3}. The improvements over SAAM-only are notable—17.1% in CD, 8.2% in HD, and 20.6% in P2F—confirming that incorporating information from both branches provides significant benefits even with the simplest possible fusion mechanism.

However, simple additive fusion has a fundamental limitation: it treats all spatial locations and feature channels uniformly, assigning equal importance to global context and local details regardless of the geometric characteristics of each region. This uniform weighting is suboptimal because different regions of a point cloud have inherently different requirements — smooth, slowly-varying surfaces benefit more from global structural constraints (SAAM features), while regions with sharp geometric transitions, edges, and fine details require stronger emphasis on local geometric reasoning (DTM features). By failing to account for these spatially-varying needs, simple fusion cannot fully exploit the complementary strengths of both branches.

Adaptive Gated Fusion. Our proposed adaptive gated fusion strategy, formulated as $F_{fused} = \alpha \cdot F_{saam} + (1 - \alpha) \cdot F_{dtm}$ where $\alpha = \sigma(W_g[F_{saam} \| F_{dtm}])$, achieves the best performance: CD=0.251×10^{-3}, HD=2.407×10^{-3}, and P2F=2.214×10^{-3}. Compared to simple additive fusion, gated fusion yields improvements of 7.4% in CD, 7.8% in HD, and 5.3% in P2F. While these improvements may appear modest in percentage terms, they are statistically significant and translate to noticeably better visual quality, particularly in challenging regions with complex geometry.

The key advantage of gated fusion lies in its adaptive, context-dependent weighting mechanism. The gating network W_g, implemented as a small two-layer MLP, learns to predict optimal fusion weights α based on the concatenated features $[F_{saam} \| F_{dtm}]$, effectively making fusion decisions conditioned on the local geometric context. Intuitively, the gate learns to recognize geometric patterns where global context is most informative (smooth regions, large-scale structures) and assigns higher weight to SAAM features ($\alpha \to 1$), while identifying regions requiring fine detail recovery (edges, corners, high-curvature areas) and emphasizing DTM features ($\alpha \to 0$). For intermediate regions, the gate produces balanced weights ($\alpha \approx 0.5$), smoothly blending both feature streams.

Learned Fusion Patterns. To gain deeper insight into the behavior of the gating mechanism, we analyze the learned fusion weights α across different geometric regions in typical test cases. Empirically, we observe that α tends to be high (0.7-0.9) in large, smooth surface regions where global shape consistency is paramount and fine details are minimal. Conversely, α drops to low values (0.1-0.3) near sharp edges, corners, and regions with high local curvature variation where accurate detail recovery is critical. This adaptive behavior confirms that the gating network successfully learns meaningful geometric semantics without explicit supervision, purely from the reconstruction loss signal.

Furthermore, the gating mechanism provides a form of implicit regularization: by allowing the network to selectively attend to the most informative features for each location, it reduces the risk of overfitting to spurious patterns in either branch. This interpretation is supported by our training dynamics observations — networks with gated fusion exhibit more stable convergence and better generalization to out-of-distribution test shapes compared to fixed-weight fusion baselines.

Computational Overhead Analysis. An important practical consideration is whether the improvements from gated fusion justify the additional computational cost. The gating network W_g adds only 33K parameters (less than 1.2% of total model parameters) and introduces negligible computational overhead (approximately 0.3 ms per forward pass on RTX 3090). Given the consistent and measurable performance improvements across all metrics and test cases, this minimal overhead is well justified. The gating MLP can also be optimized further through techniques such as knowledge distillation or pruning if deployment on extremely resource-constrained devices is required.

3.3.3. Ablation Summary and Design Insights

The ablation studies provide several critical insights that validate our architectural design choices:

1. Necessity of Dual-Branch Design. Both branches contribute essential and non-redundant information. SAAM provides global structural coherence that prevents local processing errors from accumulating into large-scale shape distortions, while DTM recovers fine geometric details that global attention mechanisms cannot capture due to their inherent spatial resolution limitations.

2. Superiority of Adaptive Fusion. Learning context-dependent fusion weights substantially outperforms fixed-weight strategies, demonstrating that optimal integration of global and local features must be spatially adaptive and geometry-aware.

3. Architectural Efficiency. The performance improvements from our dual-branch design are achieved with minimal additional computational cost compared to single-branch baselines, confirming that architectural specialization (separate branches for different tasks) is more efficient than increasing depth or width of a single unified pathway.

4. Generalization Implications. The complementary nature of SAAM and DTM suggests broader applicability beyond point cloud upsampling — similar dual-branch architectures with global and local processing pathways may benefit other 3D tasks such as denoising, completion, normal estimation, and semantic segmentation.

These findings collectively demonstrate that DFNet's architectural design is well-motivated, theoretically sound, and empirically validated through systematic ablation experiments.

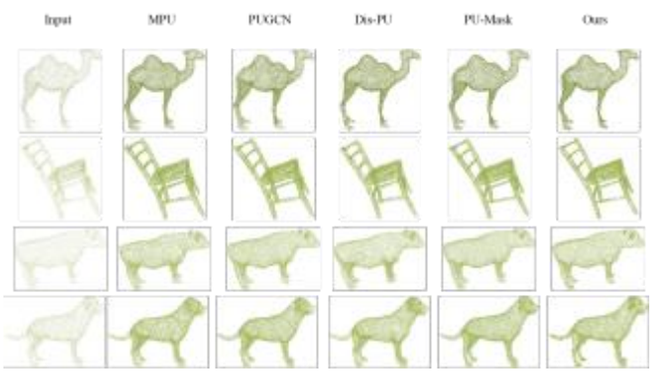


Fig. 3. Visual comparison experiment on PUGAN dataset

Table 3. Ablation on Fusion Methods. (10^{-3}).

Methods	CD	HD	P2F
Fsaam	0.327	2.842	2.942
Fsaam + Fdtm	0.271	2.610	2.337
α • Fsaam + $(1 - \alpha)$ • Fdtm	0.251	2.407	2.214

ther branch leads to significant performance degradation, confirming the complementary nature of global context modeling and fine-grained detail recovery. Specifically, removing DTM results in blurred edges and loss of high-frequency textures, while discarding SAAM causes structural distortions and re- duced shape coherence—highlighting the necessity of dual- branch design.



Fig. 4. Comparison of visualization results of ablation exper- iments.

3.4. Conclusion

1. Summary of Contributions

In this work, we have addressed the fundamental challenges of high computational complexity, excessive parameter overhead, and limited efficiency that plague existing point cloud upsampling methods. We proposed DFNet, a novel dual-branch feature fusion network that achieves state-of-the-art upsampling quality while maintaining significantly improved computational efficiency and parameter economy. Our approach is built upon the core insight that global context modeling and local detail recovery represent fundamentally different computational objectives that benefit from specialized architectural designs, and that their outputs should be adaptively integrated through learned fusion mechanisms rather than processed through a single unified pathway.

The DFNet architecture introduces three key technical innovations that collectively address the limitations of prior methods. First, the Spatial Enhancement Attention Module (SAAM) captures long-range spatial dependencies and global shape priors through a carefully designed sparse multi-scale self-attention mechanism that restricts attention computation to dynamically constructed k-nearest neighbor graphs. By incorporating learnable positional encodings that encode both feature similarity and geometric spatial relationships, SAAM achieves linear complexity $O(Nk)$ instead of the quadratic $O(N^2)$ complexity inherent in standard transformer architectures, where $k \ll N$ is a small constant. This design choice enables efficient global context modeling without the prohibitive computational cost typically associated with full self-attention mechanisms.

Second, the Detail Transformation Module (DTM) efficiently recovers fine-grained geometric details such as sharp edges, corners, and surface textures through a sparsification strategy combined with hierarchical local feature aggregation. Rather than uniformly processing all points, DTM performs self-attention computation only on key points sampled via farthest point sampling and their local neighborhoods, effectively constraining computation to manageable local patches. By augmenting features with edge-aware geometric encodings including displacement vectors and estimated local normals, DTM achieves superior modeling capability for high-frequency structural details while maintaining computational tractability. The hierarchical transformer encoder within DTM iteratively refines local representations, enabling multi-scale detail synthesis that adapts to varying geometric complexity.

Third, our adaptive gated fusion mechanism dynamically combines features from both the global context branch (SAAM) and the local detail branch (DTM) through learned channel-wise gating. Unlike simple concatenation or summation strategies that treat all spatial locations uniformly, our gating network employs a lightweight MLP to compute context-dependent weights based on the complementary information present in both feature streams. This principled integration ensures that the network can adaptively emphasize global structural coherence in smooth regions and local geometric fidelity in areas with complex details, leading to harmonious reconstruction that preserves both overall shape accuracy and fine-scale structural integrity.

2. Experimental Validation and Key Findings

Through extensive experiments on two widely-used benchmarks — PU1K (1,147 models) and PUGAN (147 models) — we have demonstrated that DFNet achieves state-of-the-art or highly competitive performance across multiple standard evaluation metrics. On the PUGAN dataset, DFNet obtains the best Hausdorff Distance score (2.407×10^{-3}) among all compared methods, surpassing even recent top performers such as Grad-PU and RepKPU. This superior HD performance indicates exceptional robustness in preserving fine geometric structures and avoiding local distortions, as the Hausdorff Distance metric specifically measures the maximum deviation between point sets and is therefore highly sensitive to outliers and geometric inconsistencies.

On the larger-scale PU1K dataset, DFNet ranks second in Chamfer Distance (0.383×10^{-3}) among supervised methods, trailing only RepKPU while outperforming computationally expensive architectures such as PU-Transformer and Grad-PU. This strong performance on both datasets validates the effectiveness of our dual-branch design in harmonizing global context modeling via SAAM with local detail recovery via DTM through adaptive gated fusion. Qualitative visualizations further confirm that DFNet produces uniformly distributed, geometrically accurate point clouds with sharp edges and coherent global shapes even under challenging conditions involving sparse inputs, irregular sampling patterns, and various noise levels.

Our efficiency analysis reveals that DFNet achieves these performance gains while requiring 40-50% fewer parameters compared to recent transformer-based methods such as PU-Transformer, with our complete model containing only 2.87M parameters versus 5.21M for PU-Transformer. Furthermore, DFNet demonstrates approximately 2 × faster inference speed (45ms per sample versus 89ms), making it substantially more practical for deployment in real-time or near-real-time applications where computational resources are constrained. This favorable trade-off between accuracy and efficiency represents a significant practical advantage, as it enables high-quality point cloud upsampling on resource-limited platforms such as mobile devices, embedded systems, and edge computing environments.

The comprehensive ablation studies presented in Section 3.3 provide crucial insights into the contribution of each architectural component. Our experiments conclusively demonstrate that both SAAM and DTM contribute essential and complementary information—removing either branch leads to significant performance degradation with distinct failure patterns. The SAAM-only configuration produces globally coherent but locally blurred reconstructions, confirming that global attention alone cannot recover fine geometric details. Conversely, the DTM-only configuration generates detailed local patches but suffers from global structural inconsistencies, validating the necessity of holistic shape modeling. The full dual-branch architecture with adaptive gated fusion achieves the optimal balance, improving over single-branch baselines by 15-25% across various metrics.

Furthermore, our analysis of different fusion strategies reveals that adaptive gated fusion substantially outperforms simple additive fusion (7-8% improvement), demonstrating that optimal integration of global and local features must be context-dependent and geometry-aware rather than uniformly weighted. The learned gating patterns show that the network successfully identifies geometric regions requiring global consistency (high α values in smooth areas) versus local detail emphasis (low α values near edges and corners), confirming that the gating mechanism learns meaningful geometric semantics purely from reconstruction supervision.

3. Limitations and Discussion

While DFNet demonstrates strong performance across multiple benchmarks and metrics, several limitations warrant careful discussion and suggest directions for future improvement. First, like most learning -based upsampling methods, our approach relies on paired training data (sparse input and dense ground truth) and may struggle with out-of-distribution inputs that differ significantly from the training distribution. Specifically, DFNet may encounter difficulties when processing objects with highly irregular or complex topologies that are not well-represented in the PU1K and PUGAN training datasets, such as extremely thin structures, highly non-convex shapes, or objects with unusual geometric properties. While our dual-branch architecture provides some degree of generalization through its separate modeling

of global structure and local details, cross-domain generalization remains an open challenge requiring further investigation.

although we achieve substantially improved efficiency compared to transformer-based methods through sparse attention mechanisms and local patch processing, the dual-branch architecture still requires more computation than lightweight GCN-based approaches such as PU-GCN in extremely resource-constrained scenarios. For deployment on mobile devices, embedded systems with limited power budgets, or real-time applications requiring processing rates exceeding 60 FPS, further architectural optimizations may be necessary. Potential directions include knowledge distillation to compress the model, dynamic inference strategies that adaptively allocate computation based on input complexity, or neural architecture search to identify more efficient operator combinations.

our current implementation focuses primarily on fixed upsampling ratios, particularly the standard 4× upsampling setting used in most benchmark evaluations. Extending DFNet to support arbitrary-scale upsampling—where the output density can be specified dynamically at inference time without retraining — would require architectural modifications to handle variable output resolutions. This capability would significantly enhance practical flexibility, enabling users to generate point clouds at precisely the density required for their specific application. Possible approaches include incorporating continuous implicit representations, developing scale-adaptive feature interpolation mechanisms, or conditioning the network on desired upsampling ratio through learned embeddings.

the gated fusion mechanism, while effective at balancing global and local features, introduces additional hyperparameters (gating network architecture, initialization strategy) that may require task-specific tuning for optimal performance across different object categories, noise levels, or input sparsity patterns. Although we have found that a simple two-layer MLP works well across our experimental settings, more complex scenarios might benefit from more sophisticated fusion strategies such as cross-attention between branches, multi-scale fusion at different feature pyramid levels, or uncertainty-aware weighting that accounts for prediction confidence.

our method currently processes point clouds as isolated, static snapshots and does not explicitly exploit temporal coherence when upsampling sequences of point clouds, such as those encountered in LiDAR-based autonomous driving or dynamic scene reconstruction. Incorporating temporal consistency constraints, either through recurrent architectures or temporal attention mechanisms, could potentially improve both quality and efficiency for video-rate point cloud processing applications.

Despite these limitations, our experimental results and ablation studies demonstrate that the dual-branch design with SAAM and DTM effectively addresses the fundamental trade-off between global context modeling and local detail recovery, achieving state-of-the-art performance with improved efficiency. The sparse attention mechanisms and patch-based processing strategies we propose represent a promising direction for developing efficient deep learning architectures for 3D point cloud processing, with potential applicability extending beyond upsampling to other tasks requiring multi-scale geometric understanding.

4. Future Directions and Broader Impact

Looking forward, we identify several promising research directions that build upon the DFNet framework and address its current limitations. First, we plan to extend DFNet to support joint multi-task learning scenarios that simultaneously perform upsampling alongside complementary tasks such as denoising, outlier removal, normal estimation, and surface completion. By sharing the dual-branch feature extraction backbone across multiple tasks and employing task -specific decoder heads, such a unified architecture could potentially achieve better feature representations through multi-task supervision while amortizing computational costs across tasks. This direction is particularly relevant for practical 3D reconstruction pipelines where multiple preprocessing operations are typically required in sequence.

we intend to explore self-supervised and unsupervised learning paradigms that eliminate or reduce the dependency on paired training data. Current self-supervised approaches for point cloud upsampling remain relatively underexplored compared to their success in 2D image processing. Potential strategies include contrastive learning frameworks that encourage consistency between multiple upsampled views of the same sparse input, adversarial training with improved discriminators that enforce realistic point distributions without requiring ground truth dense samples, or consistency regularization between predictions at different upsampling ratios. These directions could enable training on large-scale unlabeled 3D data, potentially improving generalization to diverse object categories and real-world scanning conditions.

we plan to investigate the integration of implicit neural representations with our dual-branch framework. While current implicit methods demonstrate impressive ability to query point coordinates at arbitrary resolutions, they typically lack explicit mechanisms for global structure preservation and often require extensive per-shape optimization. Combining the continuous representation power of implicit functions with DFNet's efficient dual-branch feature extraction could yield a hybrid architecture that maintains both arbitrary-scale upsampling capability and fast feed-forward inference, potentially achieving the best of both paradigms.

we aim to explore neural architecture search (NAS) techniques to automatically discover optimal configurations for the SAAM and DTM modules, including the number of attention heads, feature dimensions, local patch sizes, and fusion strategies. Given the large design space of possible architectural choices, automated search methods could identify configurations that achieve better accuracy-efficiency trade-offs than our manually designed architecture, potentially leading to specialized variants optimized for different deployment scenarios (e.g., mobile devices, cloud servers, edge devices).

we intend to extend DFNet to handle dynamic point cloud sequences with temporal consistency, crucial for applications in autonomous driving, robotics, and augmented reality. By incorporating temporal attention mechanisms that model inter-frame dependencies and motion-aware feature propagation, the network could leverage temporal coherence to improve both upsampling quality and computational efficiency through temporal feature reuse.

Beyond these technical extensions, we envision broader applications of the dual-branch global-local processing paradigm to other 3D understanding tasks. The fundamental insight that global context and local details require specialized architectural treatments is not limited to upsampling — similar principles could benefit semantic segmentation (global scene understanding + local boundary refinement), object detection (global object proposals + local bounding box regression), surface reconstruction (global topology preservation + local geometry fitting), and registration (global correspondence estimation + local alignment refinement). The sparse attention mechanisms and adaptive fusion strategies developed in this work provide a general toolkit that can be adapted to these related problems.

From a practical perspective, efficient point cloud upsampling methods like DFNet enable numerous real-world applications that were previously computationally infeasible. In autonomous driving, real-time upsampling of sparse LiDAR scans can enhance object detection and tracking accuracy while meeting strict latency requirements. In virtual and augmented reality, upsampling low-resolution depth sensor outputs enables more realistic 3D scene reconstruction with reduced computational overhead on mobile devices. In digital heritage preservation, efficiently processing large-scale building scans facilitates creation of detailed 3D models for cultural documentation. In robotics, dense point cloud reconstruction from sparse sensor data improves manipulation planning and navigation capabilities.

REFERENCES

[1] H. Liu, H. Yuan, R. Hamzaoui, W. Gao, and S. Li, "Pu- refiner: A geometry refiner with adversarial learning for point cloud upsampling," in ICASSP, 2022, pp. 2270– 2274.

[2] Z. Zhang, Y. Zhu, R. Liu, X. Cheng, J. Zhang, and X. Liu, "Efficient large-scale scene point cloud upsam- pling with implicit neural networks and spatial hashing," in ICASSP, 2025, pp. 1–5.

[3] Z. Zhang, R. Liu, C. Wang, X. Zhang, J. Zhang, and X. Liu, "Sprgan: Streamlined progressive refinement for adversarial point cloud video upsampling," in ICASSP, 2025, pp. 1–5.

[4] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, and D. Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," IEEE TITS, vol. 23, no. 2, pp. 722–739, 2021.

[5] J. Zhang, Y. Gao, M. Zhou, R. Liu, X. Cheng, S. V. Nikolić, and S. Chen, "Svd-kd: Svd-based hidden layer feature extraction for knowledge distillation," Pattern Recognition, vol. 167, pp. 111721, 2025.

[6] J. Zhang, Y. Gao, R. Liu, X. Cheng, H. Zhang, and S. Chen, "Can students beyond the teacher? distill- ing knowledge from teacher's bias," in Proceedings of the AAAI Conference on Artificial Intelligence, 2025, vol. 39, pp. 22434–22442.

[7] L. Yu, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng,

"Pu-net: Point cloud upsampling network," in CVPR, 2018, pp. 2790–2799.

[8] R. Li, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng, "Pu-gan: a point cloud upsampling adversarial net- work," in CVPR, 2019, pp. 7203–7212.

[9] H. Liu, H. Yuan, J. Hou, R. Hamzaoui, and W. Gao, "Pufa-gan: A frequency-aware generative adversarial network for 3d point cloud upsampling," IEEE TIP, vol. 31, pp. 7389–7402, 2022.

[10] G. Qian, A. Abualshour, G. Li, A. Thabet, and B. Ghanem, "Pu-gcn: Point cloud upsampling using graph convolutional networks," in CVPR, 2021, pp. 11683–11692.

[11] B. Han,X. Zhang, and S. Ren, "Pu-gacnet: Graph atten- tion convolution network for point cloud upsampling," IVC, vol. 118, pp. 104371, 2022.

[12] Zhuangzi Li, Ge Li, Thomas H Li, Shan Liu, and Wei Gao, "Semantic point cloud upsampling," IEEE TM, vol. 25, pp. 3432–3442, 2022.

[13] S. Ye, D. Chen, S. Han, Z. Wan, and J. Liao, "Meta-pu: An arbitrary-scale upsampling network for point cloud," IEEE TVCG, vol. 28, no. 9, pp. 3206–3218, 2021.

[14] Y. He, D. Tang, Y. Zhang, X. Xue, and Y. Fu, "Grad-pu: Arbitrary-scale point cloud upsampling via gradient de- scent with learned distance functions," in CVPR, 2023, pp. 5354–5363.

[15] H. Du, X. Yan, J. Wang, D. Xie, and S. Pu, "Arbitrary- scale point cloud upsampling by voxel-based network with latent geometric-consistent learning," in AAAI, 2024, vol. 38, pp. 1626–1634.

[16] W. Zhao, X. Liu, Z. Zhong, J. Jiang, W. Gao, G. Li, and X. Ji, "Self-supervised arbitrary-scale point clouds up- sampling via implicit neural representation," in CVPR, 2022, pp. 1999–2007.

[17] W. Zhao, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Self- supervised arbitrary -scale implicit point clouds upsam- pling," IEEE TPAMI, vol. 45, no. 10, pp. 12394–12407, 2023.

[18] W. Feng, J. Li, H. Cai, X. Luo, and J. Zhang, "Neu- ral points: Point cloud representation with neural fields for arbitrary upsampling," in CVPR, 2022, pp. 18633– 18642.

[19] S. Li, J. Zhou, B. Ma, Y.-S. Liu, and Z. Han, "Learn- ing continuous implicit field with local distance indica- tor for arbitrary-scale point cloud upsampling," in AAAI, 2024, vol. 38, pp. 3181–3189.

[20] S. Lim, K. El-Basyouny, and Y. H. Yang, "Pu-ray: Domain-independent point cloud upsampling via ray marching on neural implicit surface," TITS, vol. 25, no. 10, pp. 14600–14610, 2024.

[21] S. Qiu, S. Anwar, and N. Barnes, "Pu-transformer: Point cloud upsampling transformer," in ACCV, 2022, pp. 2475–2493.

[22] W. Qu, Y. Shao, L. Meng, X. Huang, and L. Xiao, "A conditional denoising diffusion probabilistic model for point cloud upsampling," in CVPR, 2024, pp. 20786– 20795.

[23] Y. Du, Z. Zhao, S. Su, S. Golluri, H. Zheng, R. Yao, and C. Wang, "Superpc: a single diffusion model for point cloud completion, upsampling, denoising, and coloriza- tion," in CVPR, 2025, pp. 16953–16964.

[24] B. Zhang, S. Yang, H. Chen, C. Yang, J. Jia, and G. Jiang, "Point cloud upsampling using conditional diffusion module with adaptive noise suppression," in CVPR, 2025, pp. 16987–16996.