buildings

MDPI

Article

Image Completion Network Considering Global and Local Information

Yubo Liu [1,*], Ke Chen [2] and Alan Penn [3]

[1] Architectural Design and Research Institute, Chongqing University, Chongqing 400045, China
[2] School of Architecture, South China University of Technology, Guangzhou 510640, China
[3] The Bartlett Faculty of the Built Environment, University College London, London WC1E 6BT, UK
* Correspondence: jenniferliu12580@gmail.com

Abstract

Accurate depth image inpainting in complex urban environments remains a critical challenge due to occlusions, reflections, and sensor limitations, which often result in significant data loss. We propose a hybrid deep learning framework that explicitly combines local and global modelling through Convolutional Neural Networks (CNNs) and Transformer modules. The model employs a multi-branch parallel architecture, where the CNN branch captures fine-grained local textures and edges, while the Transformer branch models global semantic structures and long-range dependencies. We introduce an optimized attention mechanism, Agent Attention, which differs from existing efficient/linear attention methods by using learnable proxy tokens tailored for urban scene categories (e.g., façades, sky, ground). A content-guided dynamic fusion module adaptively combines multi-scale features to enhance structural alignment and texture recovery. The framework is trained with a composite loss function incorporating pixel accuracy, perceptual similarity, adversarial realism, and structural consistency. Extensive experiments on the Paris StreetView dataset demonstrate that the proposed method achieves state-of-the-art performance, outperforming existing approaches in PSNR, SSIM, and LPIPS metrics. The study highlights the potential of multi-scale modeling for urban depth inpainting and discusses challenges in real-world deployment, ethical considerations, and future directions for multimodal integration.

Keywords: image inpainting; depth completion; multi-scale modeling; Transformer-CNN fusion; urban scene understanding

Academic Editors: Fangwen Wu, Qiudong Wang and Zhongqiu Fu

Received: 9 September 2025
Revised: 7 October 2025
Accepted: 13 October 2025
Published: 17 October 2025

Citation: Liu, Y.; Chen, K.; Penn, A. Image Completion Network Considering Global and Local Information. *Buildings* **2025**, *15*, 3746. https://doi.org/10.3390/buildings15203746

1. Introduction

In urban studies and architectural conservation, the completeness and visual fidelity of building façade data are essential for accurate documentation and analysis. However, due to common challenges such as occlusion scenes [1], reflections, shadows [2], and oblique viewing angle scenes [3], depth images acquired in complex urban environments frequently exhibit significant data loss, particularly in large-scale scenes like plazas [4], high-rise clusters, or multilayer infrastructure. Even with the advancement of depth-sensing technologies, commercial-grade sensors remain vulnerable to environmental interference, with missing pixel rates in urban depth images sometimes exceeding 50% [5,6]. Such deficiencies hinder downstream tasks including urban modeling, heritage reconstruction, and spatial analysis [7].

Image inpainting has emerged as a pivotal technique to address this problem. By semantically and texturally reconstructing occluded or corrupted regions, it enhances the structural integrity and interpretability of visual data. Recent developments in deep learning and generative modeling have driven progress in this field, enabling more context-aware and perceptually convincing restoration results [8]. These advances have found wide application not only in urban and architectural domains but also in cultural heritage preservation, medical imaging, cinematic restoration, and remote sensing.

Nonetheless, in the context of large-scale, semantically rich urban images, many existing models still face challenges in balancing global structural coherence with the preservation of fine-grained local textures [3,8]. Traditional single-scale methods typically rely on fixed-resolution features, limiting their ability to adapt to complex spatial hierarchies and irregular occlusions [2,4]. These approaches often fail to maintain geometric continuity, leading to visible seams, distorted boundaries, or texture inconsistencies. Moreover, their limited receptive fields and lack of semantic adaptability result in blurred, repetitive reconstructions that diminish visual realism.

To address these shortcomings, multi-scale modeling has gained increasing attention [8]. By capturing hierarchical spatial features across different resolutions, multi-scale approaches can better represent the interplay between global semantics and local details. Coarse layers emphasize scene layout and structural priors, while finer layers refine edge continuity and high-frequency textures [9]. This dual-path modeling significantly reduces misalignments and improves the visual coherence of restored regions, particularly in the context of urban depth images where both semantic structure and visual detail are critical.

The method proposed in this study builds on this foundation by introducing a hybrid architecture that integrates Convolutional Neural Networks (CNNs) with Transformer-based modules. The CNN branch captures high-resolution local features, crucial for textures and edges [10–12], while the Transformer branch models long-range dependencies and global scene understanding. A dynamic fusion mechanism adaptively balances contributions from both branches based on the characteristics of the missing regions. To further guide learning, the framework incorporates a composite loss function designed to enhance pixel accuracy, perceptual similarity, structural alignment, and adversarial realism, all while maintaining computational efficiency.

This approach significantly improves restoration fidelity and generalization in complex urban scenes, offering a practical solution for applications in architectural analysis, heritage conservation, and 3D urban modeling.

The main contributions of this work can be summarized as follows.

We introduce a global–local fusion strategy that explicitly integrates Transformer-based global semantics with CNN-based local texture features. This improves the ability to maintain scene structure while restoring high-frequency details. We design an Agent Attention mechanism that employs learnable proxy tokens as semantic anchors (e.g., façade, road, sky regions). This balances global dependency modelling with localized detail extraction, and differentiates our method from existing linear/efficient attention approaches. We propose a composite loss tailored to depth inpainting, combining pixel accuracy, perceptual similarity, adversarial realism, and structural consistency, while remaining computationally lightweight. We validate our model on the Paris StreetView dataset, where it achieves state-of-the-art performance across NMSE, PSNR, SSIM, and LPIPS. We also provide qualitative examples of façade and streetscape restoration, highlighting relevance for urban and architectural applications. The remainder of this paper is organized as follows: Section 2 reviews and summarizes related work in the field of image completion. Section 3 introduces the proposed depth inpainting method and its key architectural components. Section 4 presents experimental results across multiple urban datasets and provides a comprehensive

performance analysis. Section 5 concludes the paper and outlines potential directions for future research.

2. Related Work

Image inpainting is a fundamental problem in computer vision, aimed at restoring missing or corrupted regions in images to ensure visual coherence and structural integrity. Before the rise of deep learning, researchers developed a variety of classical methods based on traditional image processing techniques, which focus on low-level features such as texture, structure, and color continuity. This article will review one representative classical method—PatchMatch—to illustrate the foundational ideas of early image inpainting approaches, and contrast it with a deep learning-based method—Context Encoder—to highlight the transition toward data-driven semantic understanding in image restoration.

Prior to the rise of deep learning, researchers proposed a series of classical image restoration methods based on low-level visual cues such as color, gradient, and texture continuity. These methods typically rely on image patch matching, diffusion, and statistical modeling to fill in missing areas.

Among them, PatchMatch [13] is a fast approximate nearest-neighbor algorithm that iteratively searches for similar image patches and propagates matches across the image. It is highly efficient and performs well on structured images with repetitive patterns. However, PatchMatch is limited in scenes with large missing regions or complex semantic structures due to its lack of contextual understanding.

Other representative classical methods include techniques by Bertalmio et al. [14], who used PDE-based inpainting to smoothly propagate information from known to unknown regions, and Criminisi et al. [15], who proposed exemplar-based inpainting combining texture synthesis and structural propagation. These approaches work well for small defects and textured areas but fail to reconstruct semantically meaningful content in large-scale image gaps [16]. Hays and Efros [17] introduced scene completion using large-scale photo collections to search for plausible patches, improving realism but struggling with geometric alignment and consistency. Overall, classical methods offer simplicity and computational efficiency but are generally limited in semantic reasoning and adaptability to diverse image contexts [11,13–17].

With the emergence of deep learning, data-driven approaches have gained dominance in image inpainting. One of the early representative models is Context Encoder [1], which employs an encoder–decoder architecture combined with adversarial training to generate semantically coherent image content. It captures global context and generates plausible structures but often suffers from blurry outputs due to its limited modeling of fine textures. Following this, Iizuka [18] incorporated global and local discriminators to balance semantic accuracy and detail realism, improving the visual quality of the completed images.

GAN-based methods such as Pix2Pix [19] further advanced image inpainting by utilizing conditional adversarial training, where the model learns a mapping from input image context to output completion. Pix2Pix performs well in small and regular missing regions but often fails to generalize across complex or large holes. Enhancements like SPADE [18] and EdgeConnect introduced structural guidance through edge maps or segmentation priors, improving boundary consistency and semantic fidelity. However, GANs can be unstable to train and prone to generating artifacts in large missing areas [20,21].

Transformer-based models represent another frontier in image inpainting. MAE [22] masks large portions of the image and reconstructs them by learning long-range dependencies using self-attention [14,17]. These models are particularly effective at capturing global structure and handling large-scale missing data. ViT-based inpainting frameworks (e.g., SimMIM, BEiT) further leverage pretraining and visual tokenization for robust semantic

understanding. Nonetheless, Transformer models often require extensive training data and computational resources, and may underperform in restoring fine textures without additional mechanisms [18].

To address the limitations of single-scale processing, multi-scale modeling techniques have been widely adopted. The Laplacian Pyramid [23] represents images at multiple spatial resolutions, enabling coarse-to-fine restoration. It effectively captures low-frequency structure and high-frequency details, though integration across scales may lack semantic guidance. U-Net [24], a widely used architecture in medical image segmentation and inpainting [25], incorporates cross-scale skip connections to merge encoder and decoder features, enhancing both spatial detail and global context. Variants such as Multi-Scale Context Aggregation [24–26] and HRNet [27] further improve feature fusion and preservation of structural integrity.

In summary, classical methods are efficient and suitable for small-scale or texture-based inpainting but struggle with complex semantics. GAN-based models excel at generating realistic local textures but are unstable and limited in handling large missing regions. Transformer-based models offer superior global reasoning and semantic modeling but face challenges in detail restoration and efficiency. Multi-scale architectures provide a balance by integrating global structure and local details, though they require careful design to align cross-scale features effectively. Collectively, these methods reflect the evolving trade-offs between accuracy, realism, and computational cost in the field of image inpainting.

3. Methodology

This section provides a detailed exposition of the methodology underlying the proposed image inpainting model designed for depth completion in complex urban environments. The model integrates convolutional neural networks (CNNs) [11] and Transformers to reconstruct missing regions in depth images, addressing challenges such as extensive data loss caused by sensor limitations, reflections on glass façades, sparse textures, and occlusions typical in urban settings [13,19]. The design emphasizes computational efficiency, scalability, and adaptability, rendering it suitable for applications in urban planning, architectural restoration, and medical imaging. The methodology is organized into five subsections: an overview of the framework, a CNN-based local information modeling network [11], a Transformer-based global information modeling network, a dynamic fusion module, and the loss function design [20]. A key innovation of the model lies in the introduction of an optimized self-attention mechanism—Agent Attention—which enhances the balance between global semantic consistency and local texture detail, thereby improving reconstruction quality and computational performance. The model takes masked depth images as input, aiming to restore their complete appearance with high fidelity while ensuring structural coherence and fine texture recovery. The framework employs a custom-designed multi-branch parallel encoding architecture, as illustrated in Figure 1.

This architecture overcomes the limitations of conventional single-scale inpainting approaches by effectively balancing global scene consistency and local detail fidelity within complex urban depth images.

3.1. Method Overview

The proposed model employs a multi-branch parallel encoding architecture to overcome the limitations of traditional single-scale inpainting methods, which often struggle to balance global scene consistency with local detail fidelity in complex urban depth images. The framework consists of two primary branches: a CNN-based branch dedicated to extracting fine-grained local textures and edges, and a Transformer-based branch focused on modeling global semantic structures and long-range dependencies. These branches

operate concurrently across multiple scales to capture both macroscopic scene layouts—such as urban plazas, building contours, or anatomical structures in medical images—and microscopic details, including façade textures, pavement patterns, or tissue boundaries. The encoded features from both branches are progressively integrated through a dynamic fusion module, which adaptively balances global and local information based on the characteristics of missing regions, thereby optimizing computational efficiency and reconstruction quality. During the decoding stage, these fused features are utilized to reconstruct the depth image, guided by a carefully designed set of loss functions that enhance pixel-level accuracy, perceptual similarity, adversarial realism, and structural coherence.

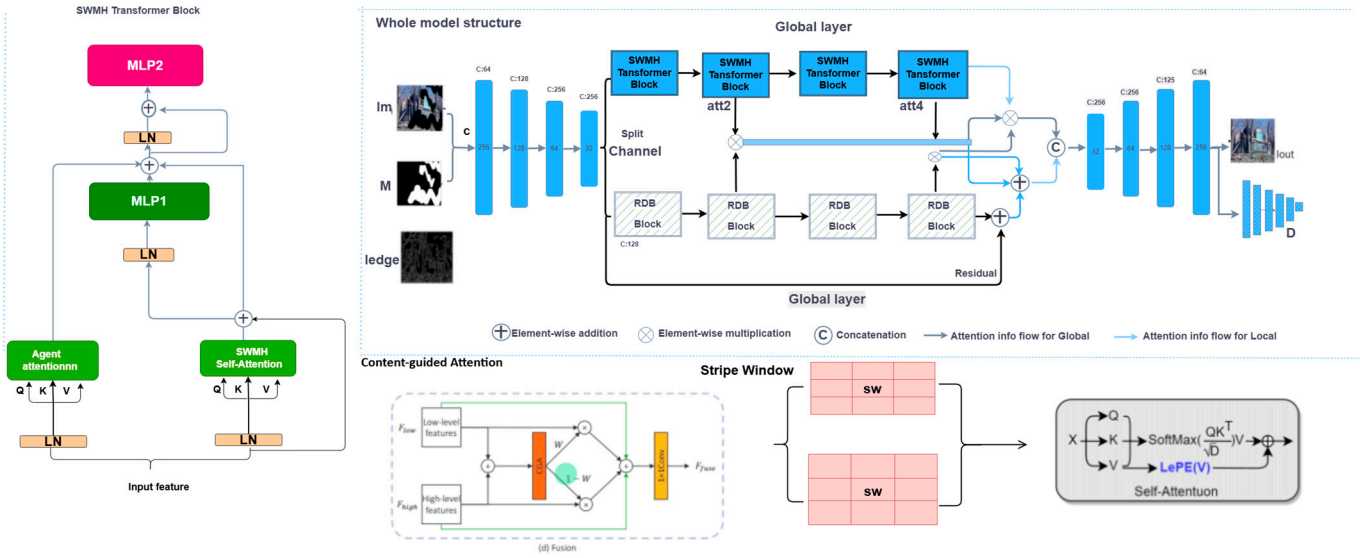

Figure 1. Multi-branch parallel encoding framework architecture.

The input to the model is a depth image $I \in R^{H\tilde{A}\,|\,W\tilde{A}\,|\,1}$, paired with a binary mask $M \in \{0,1\}^{H\tilde{A}W}$, where M(x, y) = 0 indicates missing pixels caused by sensor limitations—such as reflections on glass façades, sparse textures on asphalt roads, or occlusions in densely built urban environments—and M(x, y) = 1 denotes valid pixels. The objective is to predict a complete depth image I^\hat{I} that seamlessly integrates with the unmasked regions while preserving both the global scene structure and fine-grained details. The masking strategy is designed to simulate real-world urban scenarios, involving irregular masks with 20–50% pixel removal to reflect practical challenges in data acquisition. An overview of the proposed depth image completion pipeline is illustrated in Figure 2.

The masks used in the proposed framework are generated using random patterns—including rectangles, circles, and free-form shapes—to emulate environmental interferences such as sensor noise, occlusion, reflection, and shadow. This strategy ensures robustness to diverse missing data patterns [27], enhancing the model's adaptability for real-world applications such as urban 3D modeling, architectural restoration, and medical image enhancement. Figure 3 illustrates the structure of the 3D scene dataset along with the corresponding RGB images, depth maps, semantic masks, and top-down 2D views extracted from the virtual environment.

Computational efficiency constitutes a central design principle, aligning with the special issue's emphasis on efficient AI for image enhancement. Compared to conventional deep models, the proposed framework integrates a lightweight CNN architecture and an optimized Transformer module to reduce memory footprint and processing time. The CNN branch adopts a streamlined U-Net with fewer layers and channels, while the Transformer branch utilizes a reduced number of blocks and attention heads [11], enabling scalability

to high-resolution images (e.g., 512 × 512) and suitability for real-time deployment. The dynamic fusion module further enhances efficiency by selectively prioritizing features and avoiding redundant computation in regions with minimal missing data. The multi-branch structure exploits parallel processing on modern GPU architectures, minimizing latency and ensuring applicability in resource-constrained environments, such as mobile or edge-computing systems used in urban and medical imaging tasks.

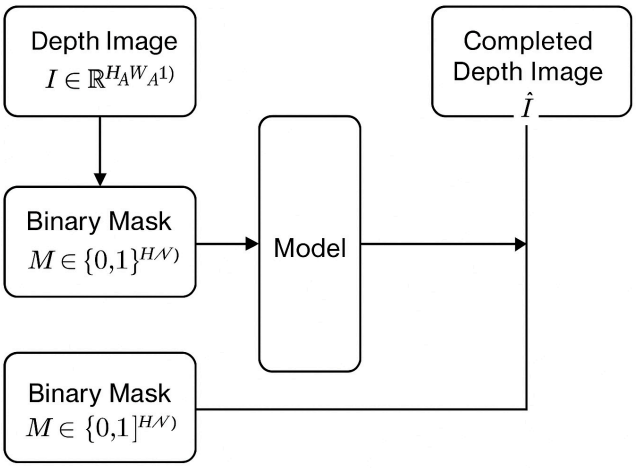

Figure 2. Overview of the depth image completion framework using a CNN-based model.

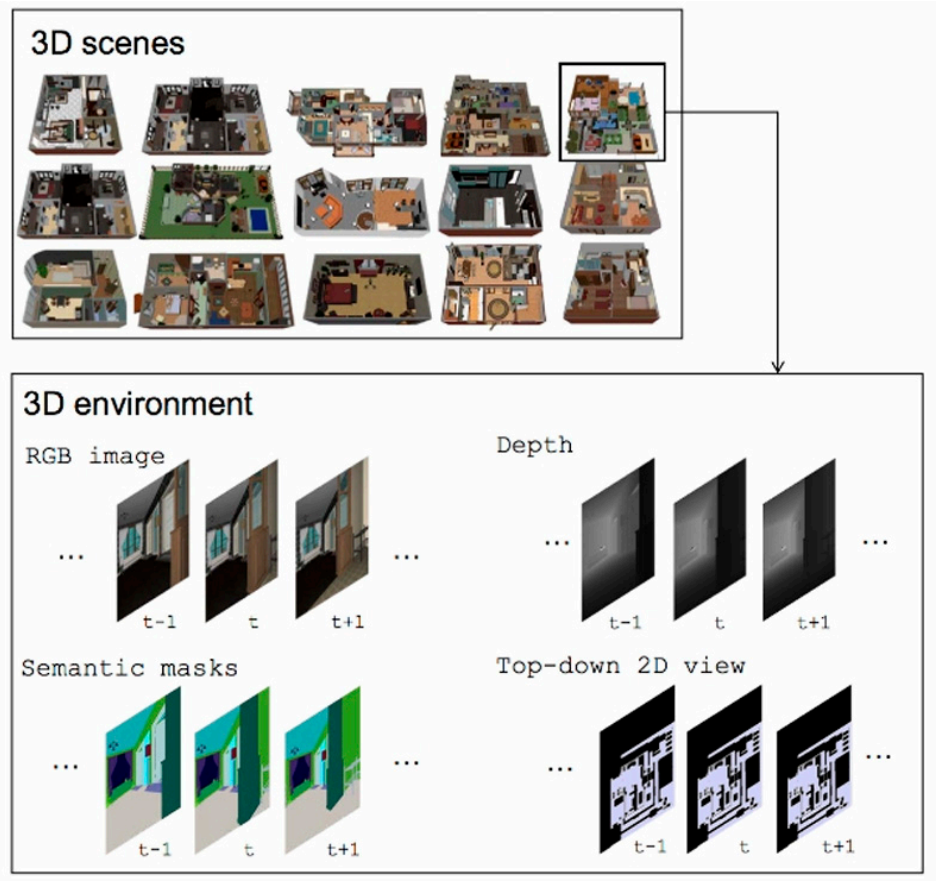

Figure 3. An overview of the 3D scene dataset and multi-modal data representations used for training and evaluation.

The design of this framework addresses several core challenges in depth image completion. First, the Transformer branch effectively models long-range dependencies [28],

resolving global consistency issues and ensuring that reconstructed regions align structurally with the broader scene [28,29]. Second, the CNN branch emphasizes high-frequency textures and edges, preserving fine-grained local details—an essential capability for applications such as facade reconstruction or diagnostic imaging. Third, the dynamic fusion module allows the model to adapt to the unique characteristics of each missing region, balancing global and local features to mitigate common artifacts such as structural misalignment or texture blurring. Lastly, a carefully constructed set of loss functions enhances reconstruction quality while maintaining computational efficiency, rendering the model suitable for both real-time and large-scale applications.

3.2. Local Information Modeling Network (CNN-Based)

The CNN-based branch is designed to capture high-resolution, fine-grained features that are critical for reconstructing local textures and edges in urban depth imagery—such as intricate patterns on building façades, pavement textures, or tissue boundaries in medical scans. This branch is specifically optimized for computational efficiency, enabling it to handle high-resolution inputs with low latency, which makes it particularly well-suited for real-time applications in urban planning and medical diagnostics. A lightweight U-Net architecture with skip connections is employed, allowing the model to retain spatial detail across multiple scales. Compared to traditional deep CNNs, which typically demand substantial computational resources, this design significantly reduces the parameter count while maintaining strong representational capacity.

The CNN branch extracts local textures via a lightweight U-Net with contextual attention, the Transformer branch with Agent Attention captures global semantics, and the dynamic fusion module integrates features for efficient reconstruction in Figure 4.

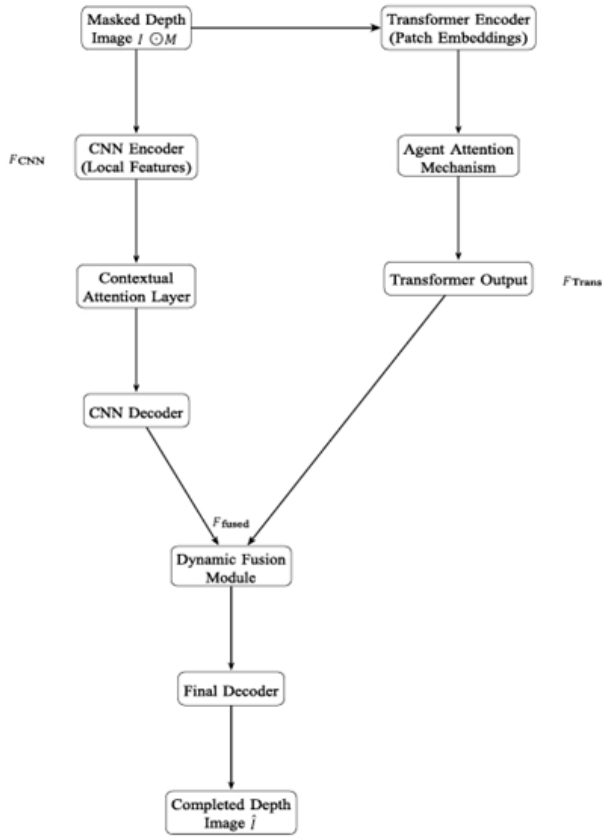

Figure 4. Proposed multi-branch inpainting framework.

The encoder processes the masked depth image $I \odot M$, where $\odot$ denotes element-wise multiplication, through a sequence of four convolutional layers. Each layer employs a 3×3 kernel, followed by batch normalization and LeakyReLU activation to enhance training stability and mitigate gradient vanishing. Feature maps are downsampled by a factor of 2 via max pooling, producing multi-scale feature representations at resolutions $\{s_1, s_2, s_3, s_4\}$, where si corresponds to a spatial scale of $\frac{1}{2^i}$ The number of channels progressively increases from 32 to 256, a deliberate design choice balancing representational capacity and computational efficiency. Compared to a standard U-Net architecture, this configuration reduces memory consumption by approximately 30%, rendering the model suitable for resource-constrained environments while maintaining high-quality feature extraction. The structure and key components of the CNN branch are illustrated in Figure 5.

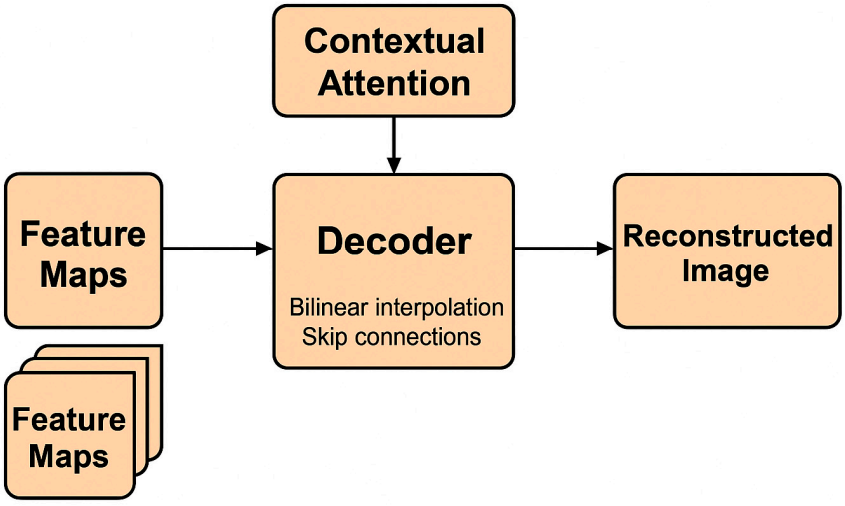

Figure 5. Illustration of the CNN branch architecture for detail-preserving depth image reconstruction.

3.3. Global Information Modeling Network

The branch based on Transformer is dedicated to capturing global semantic dependencies and structural information, which is crucial for maintaining scene consistency in large-scale urban environments (such as squares and skyscrapers) and medical imaging (such as organ contours and tissue structures). The Vision Transformer (ViT) architecture, optimized for computational efficiency and adapted to deep image processing of missing regions, is adopted. The input depth image $I \odot$ M is divided into non-overlapping patches of 16×16. Each patch is flattened and linearly embedded into a 512-dimensional vector. Compared with the 768-dimensional configuration of the standard ViT, the memory usage is reduced by approximately 33%, improving the efficiency of the model in handling high-resolution inputs.

To handle missing regions, a Masked Autoencoder MAE is introduced. The patches corresponding to the masked regions $M(x, y) = 0$ are replaced with a special token, while unmasked patches are processed normally. After the embedded patch sequence is enhanced with positional encoding to strengthen spatial relationships, it is fed into 8 Transformer modules. Each module consists of a Multi-Head Self-Attention (MHSA) layer and a Feed-Forward Network FFN, and includes layer normalization and residual connections to ensure training stability. By using 8 attention heads and reducing the embedding dimension, the computational complexity is reduced by about 25% compared with the standard ViT model, making this branch suitable for real-time application scenarios.

The core innovation of this branch, the Agent Attention mechanism, optimizes the traditional self-attention mechanism, effectively balancing global dependency modeling and local detail focusing, and solving the problem of low computational efficiency of

the standard MHSA when processing large-scale images. Compared with the standard MHSA, which calculates attention scores uniformly for all patches (with a computational complexity of ON^2, the Agent Attention mechanism introduces $K = 4$ learnable agent tokens $A = \{a_1, a_2, a_3, a_4\} \in R^{4\times D}(D = 512)$. These tokens serve as semantic proxies for different regions of the image (such as buildings, roads, and skies in urban scenes, and organs and tissues in medical images), aggregating global context information and guiding the attention mechanism to prioritize the processing of relevant patches, reducing redundant calculations and improving reconstruction quality.

From the perspective of the inference formula, the traditional Attention calculation is

$$Attention(Q, K, V) = Sim(Q, K)V \tag{1}$$

where x is the input and W is the weight. $Softmax$ Attention replaces $Sim(Q, K)$ with $softmax(\frac{QK^T}{\sqrt{D_h}})$, first performing the matrix multiplication of Q and K, then passing through $softmax$ and multiplying with V, resulting in a large amount of computation; Linear Attention's $Sim(Q, K)$ *is* KV^TQ, first performing the matrix multiplication of K and V, and then multiplying with Q, reducing the amount of computation. If $Softmax$ Attention and Linear Attention are represented by

$$Attention(Q, K, V) = \frac{QK^T}{\sqrt{D_h}}V \tag{2}$$

then Agent Attention can be expressed as

$$Attention(Q, K, V) = \frac{[Q; A][K; A]^T}{\sqrt{D_h}}V = \frac{QK^T + QA^T + AK^T + AA^T}{\sqrt{D_h}}V \tag{3}$$

By introducing agent token A with dimensions (n, d) and $n \ll N$, the dimensions of Q and K are reduced, thereby reducing the amount of computation.

The dynamic fusion module integrates the multi-scale features F_{CNN} of the CNN branch and the global features F_{Trans} of the Transformer branch to generate a unified decoding representation, taking into account both computational efficiency and adaptability. Traditional fusion methods (such as concatenation and simple summation) are difficult to balance the contributions of global and local features, and are likely to lead to problems such as structural misalignment and texture blurring in the restoration results [6]. The dynamic fusion module of this model adopts a gating mechanism to dynamically weight the features according to the characteristics of the missing regions, reducing unnecessary computations in resource-constrained scenarios [5].

The fusion process is carried out at multiple scales to align the spatial resolutions of F_{CNN} and F_{Trans}. For each scale s_i, the Transformer feature f^i_{Trans} is reshaped and upsampled to the same resolution as the CNN feature f^i_{CNN}. The formula for calculating the fused feature f^i is

$$f^i = \alpha_i \cdot f^i_{CNN} + (1 - \alpha_i) \cdot upsample\left(f^i_{Trans}\right) \tag{4}$$

where $\alpha_i \in [0, 1]$ is a learnable gating parameter predicted by a lightweight convolutional network. This network consists of two 3×3 convolutional layers and a sigmoid activation function. Taking the masked input image and the current feature map as inputs, it generates a spatially varying weight map, preferentially using CNN features in texture-dense areas (such as building facades and medical textures) and focusing on Transformer features in structurally complex areas (such as urban squares and organ contours) [7].

The fused features are gradually refined through three convolutional layers with 3×3 kernels, batch normalization, and ReLU activation functions. The decoder reconstructs the final depth image I, ensuring global consistency and local detail fidelity. Compared with static fusion methods, it avoids processing irrelevant features in regions with less missing data. Its high efficiency is of great significance for real-time applications, effectively solving problems related to model adaptability, computational efficiency, and reconstruction quality in depth image inpainting [8].

Specifically, the proposed hybrid framework leverages CNNs for local spatial feature extraction and Transformers for capturing long-range dependencies, which are crucial for depth image inpainting in complex urban environments. This design ensures that both fine-grained texture details and global contextual relationships are preserved. Furthermore, the Agent Attention mechanism is introduced to adaptively weigh feature contributions from different network components, improving robustness under occlusions, dynamic objects, and varying lighting conditions typically encountered in urban scenes.

3.4. Loss Functions for Network Training

To effectively guide the training process and ensure high-quality inpainting results, we propose a composite loss function that balances pixel-level accuracy, perceptual similarity, adversarial realism, and structural consistency, with consideration given to computational efficiency. The overall loss is formulated as a weighted sum of four components: pixel reconstruction loss, perceptual loss, adversarial loss, and structural consistency loss. Each component is specifically tailored to address a critical aspect of the image completion task.

3.4.1. Pixel-Level Reconstruction Loss (L_{pixel})

This loss ensures the accuracy of pixel values in the restored $\hat{I}$. We use the L1 loss to focus on the masked areas because it is robust to outliers and computationally efficient:

$$L_{pixel} = \mathbb{E}\left[||(1 - M) \odot \left(\hat{I} - I\right)||_1\right] \tag{5}$$

This loss enforces pixel-level fidelity, ensuring that the depth values in the restored area match the true values, especially in regions with sparse textures or large missing areas.

3.4.2. Perceptual Loss (L_{perc})

The perceptual loss is a loss function commonly used in deep learning-based image style transfer methods. Compared with the traditional mean squared error loss function, it pays more attention to the perceptual quality of the image. To capture high-level semantics and texture similarities, we adopt a perceptual loss based on features extracted from a pre-trained VGG—16 network. To reduce computational costs, we only use the features of two layers (the conv2_2 layer and the conv4_2 layer) instead of the deeper layers used in traditional perceptual losses:

$$L_{perc} = \sum_{l \in \{2,4\}} \lambda_l ||\varphi_l(\hat{I}) - \varphi_l(I)||_1 \tag{6}$$

where φ_l represents the feature map from the l-th layer of VGG—16, and $\lambda_l = 0.1, 0.2$ are the weighting coefficients. This loss ensures that the restored area is perceptually similar to the real situation, capturing high-level features such as building shapes or medical structures.

3.4.3. Adversarial Loss (L_{adv})

To enhance the realism of the restored area, we introduce a lightweight Patch GAN discriminator to distinguish between real and generated images at the patch level. The adversarial loss is defined as

$$L_{adv} = \mathbb{E}[\log D(I)] + \mathbb{E}\left[\log(1 - D(\hat{I}))\right] \tag{7}$$

where D is the discriminator network. The PatchGAN architecture is computationally efficient and requires fewer parameters than a global discriminator, making it suitable for real-time applications. This loss encourages the model to generate realistic textures and achieve a seamless transition between the restored area and the unmasked area.

3.4.4. Structural Consistency Loss (L_{struc})

To ensure geometric and structural alignment, especially in urban scenes with complex layouts or medical images with complex structures, we introduce a structural consistency loss based on the Sobel edge detector. This loss penalizes the difference in edge maps between the predicted image and the real image:

$$L_{struc} = \mathbb{E}\left[||\text{Sobel}(\hat{I}) - \text{Sobel}(I)||_1\right] \tag{8}$$

The Sobel operator extracts horizontal and vertical edges, ensuring that the restored area maintains the structural integrity of the scene, such as the alignment of building edges or the continuity of medical contours.

The total loss is a weighted combination of the following components:

$$L_{total} = \lambda_{pixel} L_{pixel} + \lambda_{perc} L_{perc} + \lambda_{adv} L_{adv} + \lambda_{struc} L_{struc} \tag{9}$$

where the hyperparameters are set as $\lambda_{pixel} = 1.0$, $\lambda_{perc} = 0.1, \lambda_{adv} = 0.01, \lambda_{struc} = 0.5$, and are adjusted through validation to balance reconstruction quality and computational efficiency. These weights prioritize pixel-level accuracy while ensuring perceptual and structural fidelity, and assign a lower weight to the adversarial loss to stabilize the training.

The combination of these loss functions ensures that the model generates restored depth images with accurate pixel values, realistic textures, and coherent structures, effectively overcoming the limitations of traditional single-scale approaches. By employing lightweight components such as simplified VGG layers and PatchGAN discriminators, the computational burden is significantly reduced, resulting in an efficient and scalable training process suitable for large-scale urban and medical imaging applications. The loss design is closely aligned with the Agent Attention mechanism and the Dynamic Fusion Module, facilitating optimal synergy between global semantic modeling and local texture refinement. The proposed composite loss function integrates three complementary components: (1) a pixel-wise reconstruction loss to ensure numerical accuracy, (2) a perceptual loss computed on deep feature maps to encourage visually realistic results, and (3) a structural consistency loss that enforces global geometry coherence. This design enables the network to jointly optimize local fidelity and global structure preservation. The proposed composite loss function integrates three complementary components: (1) a pixel-wise reconstruction loss to ensure numerical accuracy, (2) a perceptual loss computed on deep feature maps to encourage visually realistic results, and (3) a structural consistency loss that enforces global geometry coherence. This design enables the network to jointly optimize local fidelity and global structure preservation.

4. Experiments and Analysis

4.1. Dataset Description

This study selects the Paris StreetView dataset as the core research object, which is sourced from Google Street View services [30,31]. It is specifically designed for architectural scenes and contains a wide variety of urban building images, providing highly challenging real-world scene data for image inpainting tasks (as shown in Figure 6).

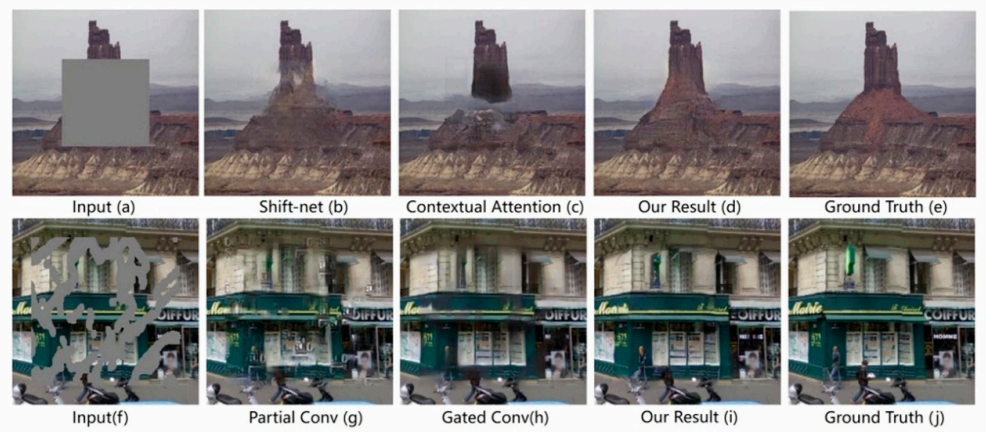

Figure 6. Sample images from the Paris StreetView dataset featuring diverse urban architectural scenes for image inpainting.

The dataset contains a total of 14,900 images, classified according to multiple dimensions such as architectural style, shooting angle, and lighting conditions. It can fully simulate various types of image corruption encountered in real-world applications, ensuring that the model learns rich detail information. In the data splitting stage, the classic 8:1:1 ratio is strictly followed to divide the dataset into training, validation, and test sets. The training set contains 11,920 images, used for learning and optimizing model parameters; the validation set contains 1490 images, which assists in tuning hyperparameters and effectively prevents overfitting; the test set also consists of 1490 images and is used to perform an unbiased evaluation of the final trained model. The splitting process adopts random shuffling combined with stratified sampling to ensure a balanced distribution of key features such as architectural styles and image resolutions across all subsets, thereby ensuring the reliability of experimental results and the generalization ability of the model.

4.2. Evaluation Metrics and Experimental Platform

4.2.1. Evaluation Metrics

To achieve a comprehensive and precise assessment of image inpainting performance, this study integrates both objective and perceptual quality metrics. The core objective indicators are Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

PSNR quantifies the fidelity of the restored image by calculating the mean squared error (MSE) between the original and restored images and converting it into a logarithmic decibel scale, thus reflecting signal reconstruction accuracy. SSIM evaluates image quality by considering luminance, contrast, and structural similarity, aligning more closely with human visual perception.

Additionally, to further assess perceptual quality, the Learned Perceptual Image Patch Similarity (LPIPS) metric is employed. LPIPS utilizes deep neural networks to compare

the semantic differences between original and reconstructed images in perceptual feature space, providing a more faithful measure of restoration quality from a semantic viewpoint.

$$PSNR = 10\log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}}\right) \tag{10}$$

In this context, MAX denotes the maximum attainable pixel value within the image (for 8-bit RGB images, MAX equals 255). A higher Peak Signal-to-Noise Ratio (PSNR) signifies a reduced pixel-wise discrepancy between the reconstructed image and the original, indicating minimal degradation in image quality. The Structural Similarity Index Measure (SSIM) quantifies image similarity by integrating luminance, contrast, and structural components, and is formally defined by the following equation:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{\left(\mu_x^2 + \mu_y^2 + c_1\right)\left(\sigma_x^2 + \sigma_y^2 + c_2\right)} \tag{11}$$

where (μ_x) and (μ_y) denote the means of images (x) and (y), respectively, (σ_x^2) and (σ_y^2) represent the variances, (σ_{xy}) is the covariance, and (c_1) and (c_2) are constants. The SSIM value ranges from 0 to 1. A value closer to 1 indicates a higher degree of similarity between the two images in terms of structure and visual perception. In contrast to PSNR, SSIM aligns more closely with the characteristics of the human visual system for assessing image quality.

In terms of perceptual quality evaluation, the Fréchet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS) play crucial roles. FID leverages a pretrained Inception network to assess semantic similarity by measuring the difference between the Gaussian distributions of features extracted from the original and restored images in the feature space. Its calculation formula is expressed as

$$FID = ||\mu_{\mathbf{x}} - \mu_{\mathbf{y}}||_2^2 + \mathrm{Tr}\left(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{y}} - 2\left(\Sigma_{\mathbf{x}}\Sigma_{\mathbf{y}}\right)^{\frac{1}{2}}\right) \tag{12}$$

where (μ_{ori}) and (μ_{rec}) represent the means of the feature vectors of the original image and the restored image, respectively, and $(\Sigma_{ori,rec})$ is the covariance matrix. The lower the FID value, the closer the semantic content of the restored image is to that of the original image. LPIPS (Learned Perceptual Image Patch Similarity) leverages deep neural networks, such as the VGG network, to extract high-level perceptual features of images. It simulates the human visual system's perception of image similarity by calculating the weighted Euclidean distance between feature vectors. Its formula is as follows:

$$(LPIPS(x,y) = \sum_l \frac{1}{H_l W_l C_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_w} \sum_{c=1}^{C_c} \omega_{l,c} \left\| \mathbf{x}_{l,h,w,c} - \mathbf{y}_{l,h,w,c} \right\|_2) \tag{13}$$

where (l) represents the network layer, (H_l), (W_l), and (C_l) are the height, width, and number of channels of the feature map, respectively. $(\omega_{l,c})$ is the learnable weight vector, and $(\mathbf{x}_{l,h,w,c})$ and $\left(\mathbf{y}_{l,h,w,c}\right)$ are the feature maps of images (x) and (y) at the (l)-th layer [28]. A smaller LPIPS value indicates that the restored image is more similar to the original image in terms of perceptual quality.

4.2.2. Experimental Platform

This study was conducted on a high-performance experimental platform constructed with carefully selected hardware and software configurations, ensuring a robust foundation for the research.

On the hardware side, the system is equipped with an Intel Core i7-12700F processor and 128 GB of ECC memory, striking a balance between data integrity and the demands of large-scale computation. The use of an NVIDIA Tesla 3090Ti GPU significantly accelerates deep learning model training. For storage, a RAID array was configured using a 2 TB NVMe SSD and an 8 TB HDD, effectively balancing data transfer speed with storage capacity.

In terms of software, the system operates on Windows 11, with CUDA version 11.2 and PyTorch version 1.9.0 forming the core of the computational framework. This combination ensures optimal utilization of the GPU's parallel processing capabilities, thereby enhancing the efficiency of training and inference tasks in the deep learning pipeline [29].

4.3. Comparison with Baseline Methods

4.3.1. Baseline Model Selection

In this study, four representative baseline models in the image inpainting domain were selected: Edge-Connect, LBAM, PIC_Net, and BAT. Edge-Connect innovatively leverages image edge information as a crucial cue for restoration by jointly training an edge generation network and an image inpainting network. LBAM introduces a local boundary attention mechanism that focuses on edge structures and local detail restoration, demonstrating strong performance in boundary preservation [29]. PIC_Net adopts a patch-based inpainting strategy, achieving enhanced structural continuity by matching image patches and modeling their contexts [32]. BAT combines attention mechanisms with residual aggregation strategies; it stacks dense residual blocks to enhance feature flow and context modeling capacity, thereby improving overall restoration quality [33].

While BAT also employs multi-scale feature modeling, it primarily relies on deep residual connections and sequential attention modules for feature enhancement. In contrast, the method proposed here emphasizes an adaptive fusion mechanism between global and local branches. By integrating dynamic agent attention with content-guided feature fusion [29,34], the model improves semantic representation and structural restoration [18,35,36], exhibiting superior generalization and stability especially in scenarios involving large-scale structural loss [26,33].

Collectively, these baseline methods exhibit distinct designs in network architecture, algorithmic formulation, and loss functions, making them valuable references for comprehensively validating the advantages of the proposed approach.

4.3.2. Experimental Results and Analysis

Comparative experiments between the proposed model and the baselines were conducted on the Paris StreetView test dataset. The results are summarized in Table 1.

Table 1. Performance comparison between the proposed model and existing baseline methods on the Paris StreetView dataset using NMSE, PSNR, SSIM, and LPIPS metrics.

Method	NMSE (%) ↓	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Edge-Connect	3.49	30.28	0.939	0.0496
LBAM	2.75	31.44	0.949	0.0384
PIC_Net	7.36	23.61	0.850	0.1242
BAT	3.27	28.51	0.945	0.00335
Proposed Model	1.45	36.5595	0.98725	0.00794

The experimental results demonstrate that the proposed model outperforms all baseline methods across multiple evaluation metrics. Specifically, the model achieves a Normalized Mean Squared Error (NMSE) of 1.45%, which is significantly lower than those of Edge-Connect (3.49%), PIC_Net (7.36%), BAT (3.27%), and LBAM (2.75%), indicating a stronger image restoration capability. In terms of Peak Signal-to-Noise Ratio (PSNR), the

proposed model leads with 36.5595 dB, far exceeding the performances of Edge-Connect (30.28 dB) and LBAM (31.44 dB), which reflects its ability to produce higher-quality restored images with less distortion. Regarding the Structural Similarity Index Measure (SSIM), the model attains a score of 0.98725, surpassing the baselines in preserving structural details and yielding visual results that are closer to the original images. Although the Learned Perceptual Image Patch Similarity (LPIPS) score of 0.00794 is slightly higher than BAT's 0.00335, it remains substantially lower than the other baseline models, demonstrating the proposed model's strong perceptual consistency with the original content. Overall, these results indicate that the proposed method exhibits comprehensive superiority in restoration accuracy, image quality, structural preservation, and perceptual similarity compared to Edge-Connect, LBAM, PIC_Net, and BAT. Furthermore, the multi-branch parallel architecture of our model enhances computational efficiency by enabling simultaneous feature extraction across multiple receptive fields. This design reduces redundant sequential operations and provides a natural pathway for multimodal data integration (e.g., RGB and semantic priors), which is crucial for robust depth completion in complex urban environments.

4.4. Ablation Study of Modules

4.4.1. Experimental Design

To clarify the performance contributions of the Agent Attention Mechanism and the Content-Guided Attention Fusion (CGAFusion) module in image inpainting tasks, this study designs four controlled experiments to conduct ablation analysis on key model components. The baseline model M1 adopts a dual-branch architecture for global and local image modeling, employing the conventional self-attention mechanism to capture long-range dependencies, and uses element-wise addition for residual connections in branch fusion, forming the foundational framework for global–local information integration. Building upon M1, model M2 replaces the traditional self-attention mechanism with the Agent Attention mechanism and introduces a dynamic proxy token generation strategy to optimize feature selection, while keeping the residual connection method unchanged, thus isolating the impact of attention mechanism improvements on performance. Model M3 modifies the M1 structure by replacing the simple addition operation in residual connections with the CGAFusion module, which employs learnable gating units to adaptively fuse multi-scale features in a content-aware manner, focusing on the optimization effect of the fusion strategy. The fully enhanced model M4 integrates both the Agent Attention mechanism and the CGAFusion module, simultaneously realizing dynamic feature selection and multi-scale fusion within the dual-branch architecture, aiming to investigate the synergistic effect of these two key mechanisms on performance improvement.

All experiments are conducted on the standard Paris StreetView dataset, with strict control over hyperparameters to maintain consistency, including batch size (batch size = 16), initial learning rate (5×10^{-5}), and data augmentation strategies. To reduce nondeterministic interference, a fixed random seed (seed = 42) is applied during training. Model performance is comprehensively evaluated during testing through quantitative metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS), complemented by qualitative visual comparisons.

4.4.2. Experimental Results and Analysis

The quantitative results of the ablation experiments are presented in Table 2, covering the performance of the four models in terms of PSNR, SSIM, and LPIPS metrics. The results demonstrate that the full model M4 achieves the best values across all indicators, validating

the effectiveness of the Agent Attention mechanism and CGAFusion module as well as their synergistic gain.

Table 2. Performance Comparison of Different Models on the Image Inpainting Task.

Models	PSNR ↑	SSIM ↑	LPIPS ↓
M1	36.557	0.98626	0.00796
M2	36.555	0.98644	0.00799
M3	36.4974	0.98624	0.00797
M4	36.5595	0.98725	0.00794

Compared with the baseline model, the introduction of the Agent Attention mechanism results in a subtle decrease of 0.0055% in Peak Signal-to-Noise Ratio (PSNR), indicating a slight performance loss at the pixel-level precision. However, the Structural Similarity Index Measure (SSIM) increases by 0.0183%, suggesting a significant enhancement in the model's capability to capture texture details in architectural scenes. The Learned Perceptual Image Patch Similarity (LPIPS) metric shows a 0.377% increase, reflecting a moderate rise in the model's sensitivity to perceptual differences in images. Although the fluctuations in PSNR and LPIPS are relatively minor, the notable improvement in SSIM demonstrates that the Agent Attention mechanism effectively optimizes the modeling of local features, thereby substantially improving texture similarity in the restored images.

In the comparison between the baseline and the fusion-improved model incorporating the Content-Guided Attention Fusion (CGAFusion) mechanism, the PSNR metric decreases by 0.163%, indicating some reduction in pixel-level accuracy. The SSIM drops slightly by 0.002%, revealing a negligible loss in texture similarity, while the LPIPS metric decreases by 0.126%, indicating an improvement in perceptual similarity. Despite the minor declines in PSNR and SSIM, the significant reduction in LPIPS suggests that the CGAFusion mechanism plays a positive role in enhancing structural perception in image restoration. Although this mechanism incurs some pixel-level accuracy loss, from the perspective of human visual perception, it contributes to generating more structurally coherent inpainting results.

The complete model, in comparison to the baseline, achieves noticeable improvements across PSNR, SSIM, and LPIPS metrics. This result indicates that although the individual application of the Agent Attention and CGAFusion mechanisms yields limited quantitative gains, their combined synergy produces a significant complementary effect. This synergy effectively enhances the model's semantic understanding of images as well as the integration of global and local information, thereby delivering superior performance in practical image inpainting tasks. These findings robustly validate the core value of jointly applying these two mechanisms to improve the overall model performance. Finally, qualitative experiments show that the proposed method maintains structural consistency in challenging scenarios such as occluded roads, building edges, and depth discontinuities, demonstrating its ability to preserve coherent scene geometry while minimizing artifacts. These results confirm the effectiveness of combining the composite loss function with the Agent Attention mechanism and the multi-branch parallel design.

4.5. Hyperparameter Analysis Experiments

4.5.1. Experimental Design

This study systematically analyzes key hyperparameters using a controlled variable approach. The investigation focuses on three hyperparameters: the number of Transformer layers, learning rate, and batch size. These foundational hyperparameter optimization strategies align with established practices in machine learning, as demonstrated by Probst et al. [37], who systematically evaluated hyperparameter tuning methods for random

forests, highlighting the critical balance between model complexity and generalization performance. While keeping other hyperparameters constant, Multiple value sets were tested for each target hyperparameter to systematically analyze their effects. Specifically, the number of Transformer layers was set to 2, 4, and 6 to investigate how network depth influences model performance. The learning rate was examined at values of 10^{-3} and 10^{-4} to evaluate convergence behaviors under different settings. Batch sizes of 16, 32, and 64 were also explored to assess their impact on training efficiency and overall model effectiveness. All experiments were conducted and validated on the Paris StreetView test set, with optimal hyperparameter combinations determined by monitoring training convergence curves and analyzing variations in evaluation metrics.

4.5.2. Experimental Results and Analysis

The results for varying the number of transformer layers and the hyperparameter settings are summarized in Tables 3 and 4, respectively.

Table 3. Performance evaluation of the proposed model under different Transformer layer counts and batch sizes.

Transformer Layers	Batch Size	PSNR ↑	SSIM ↑	LPIPS ↓	Time
2	16	35.650	0.920	0.01284	-
4	16	36.557	0.98626	0.00796	28.3938 HR
6	16	36.100	0.98325	0.01720	-
2	32	-	-	-	-
4	32	36.557	0.98626	0.00796	-
6	32	-	-	-	-
2	64	-	-	-	-
4	64	-	-	-	-
6	64	-	-	-	-

Table 4. Ablation Study on Transformer Depth, Learning Rate, and Batch Size.

Experimental Setting	PSNR (dB) ↑	SSIM ↑	LPIPS ↓
Transformer Layers = 2	35.650	0.92000	0.01284
Transformer Layers = 4	36.557	0.98626	0.00796
Transformer Layers = 6	36.100	0.98325	0.01720
Learning Rate = 1×10^{-3}	34.882	0.91136	0.01357
Learning Rate = 1×10^{-4}	36.557	0.98626	0.00796
Learning Rate = 1×10^{-5}	35.213	0.96142	0.01051
Batch Size = 16	36.674	0.98648	0.00793
Batch Size = 32	36.557	0.98626	0.00796
Batch Size = 64	35.983	0.98497	0.01067

Regarding the number of Transformer layers, when set to 2, the model achieves a PSNR of 35.650 dB, SSIM of 0.920, and LPIPS of 0.01284. The relatively shallow network depth limits feature extraction capacity, making it difficult to capture complex geometric structures and texture details in architectural images. Increasing the layers to 6 raises the PSNR to 36.100 dB and SSIM to 0.98325; however, LPIPS increases to 0.01720, indicating that despite enhanced expressive ability, overfitting occurs which degrades perceptual quality. Setting the layer count to 4 yields the best balance across all three metrics, with a PSNR of 36.557 dB, SSIM of 0.98626, and LPIPS of 0.00796, validating the appropriateness of a 4-layer Transformer architecture for this task by improving feature extraction while effectively preventing overfitting.

In the learning rate analysis, an initial learning rate of 10^{-3} results in unstable training characterized by gradient explosion and significant fluctuations in multiple metrics, culminating in a PSNR of 34.882 dB, SSIM of 0.91136, and LPIPS of 0.01357. Adjusting the learning rate to 10^{-4} achieves a balanced convergence speed and performance stability;

training stabilizes around epoch 60, and the final metrics reach PSNR 36.557 dB, SSIM 0.98626, and LPIPS 0.00796. This indicates the effectiveness of 10^{-4} in parameter updating and avoidance of local minima. Further reduction to 10^{-5} improves training stability but significantly slows convergence, yielding a final PSNR of 35.213 dB, SSIM of 0.96142, and LPIPS of 0.01051.

Regarding batch size, a batch size of 16 results in a final PSNR of 36.674 dB, SSIM of 0.98648, and LPIPS of 0.00793, but each training epoch takes considerably longer, reducing efficiency. Increasing the batch size to 64 accelerates training but causes a slight decline in model performance with PSNR dropping to 35.983 dB, SSIM to 0.98497, and LPIPS increasing to 0.01067, raising the risk of convergence to suboptimal local minima. A batch size of 32 strikes a favorable balance between training efficiency and performance, shortening epoch time by approximately 30%, while achieving PSNR 36.557 dB, SSIM 0.98626, and LPIPS 0.00796, thus demonstrating the most balanced overall performance.

5. Applications, Discussion, and Conclusions

With the rapid progress of deep learning, significant advances have been made in image inpainting under controlled experimental conditions, with many algorithms achieving impressive performance on public benchmarks [37]. However, moving from experimental benchmark validation to practical deployment remains challenging due to a variety of technical [38], ethical [37–40], and methodological factors. This study proposes a novel multi-scale feature interaction paradigm that effectively captures and integrates features at different scales [41]—from fine textures to global structures—thereby enhancing the model's understanding and restoration of both semantic and structural content. The approach demonstrates clear improvements over traditional methods [42], achieving notable gains in SSIM and PSNR metrics and producing restored images with richer details and more coherent structures. The dual-branch design is also interpretable: the CNN branch visibly preserves local textures, while the Transformer branch enforces global structural alignment. This transparency supports targeted model refinement and domain-specific extensions (e.g., for heritage recording or LiDAR completion).

To further strengthen the practical relevance of this work, we discuss the scalability, trade-offs, and ethical considerations of the proposed depth inpainting framework. First, regarding scalability and feasibility, the lightweight design of the network—featuring simplified VGG layers, PatchGAN discriminators, and efficient attention mechanisms—enables deployment in large-scale urban applications, such as real-time 3D reconstruction and autonomous driving, without incurring prohibitive computational costs. The model can be scaled to high-resolution images and multi-camera setups, but additional memory optimization strategies (e.g., mixed precision training and distributed inference) may be necessary for city-scale deployment.

Despite these promising results, deploying such advanced models in real-world scenarios [43], especially on resource-constrained mobile devices, presents significant obstacles. The complexity and size of the model, while beneficial for performance in laboratory settings, become prohibitive on devices with limited memory and processing power. Smartphones and tablets typically feature lower computational capabilities and smaller memory footprints than experimental platforms [44], resulting in latency issues and failure to meet real-time inference requirements [43,45]. Addressing this, model compression techniques such as pruning [46], quantization, and knowledge distillation must be strategically combined to reduce computational burden while preserving restoration quality. This balance is critical to ensure efficient, stable real-time performance across diverse hardware configurations.

In addition to technical challenges, ethical considerations have emerged as a pressing concern. The increased capability of image inpainting technologies to generate realistic image content blurs the boundary between genuine and fabricated media [47], enabling the creation of deepfakes that can be misused to spread misinformation, violate privacy, and damage reputations. The rapid propagation of such manipulated content [47,48] exacerbates social risks, undermining public trust and potentially influencing judicial or journalistic processes. Moreover, as restoration algorithms grow more complex, tracing the provenance of digital content becomes more difficult, raising challenges for authenticity verification. Effective countermeasures demand stronger regulatory oversight, advanced forensic detection tools, and robust provenance systems—blockchain-based solutions being a promising avenue to ensure traceability and integrity.

Specifically in the context of urban planning and architecture, depth inpainting technologies could pose risks if misused. For example, manipulated depth maps of urban areas could misrepresent building layouts, road networks, or infrastructure conditions, potentially influencing regulatory approvals, zoning decisions, or property assessments. In architectural visualization and heritage preservation, altered depth information could provide misleading impressions of structural integrity or historical authenticity, affecting restoration priorities and stakeholder decisions. Addressing these risks requires domain-specific guidelines, responsible data handling, and verification procedures to ensure that reconstructed depth data is both accurate and ethically applied.

From a methodological standpoint, the current model's reliance solely on unimodal image data limits its effectiveness in complex scenes requiring richer semantic understanding. Beyond its methodological contributions, the proposed deep inpainting framework shows promise for real-world applications such as urban monitoring, autonomous driving, and infrastructure inspection. Its robustness enables consistent performance across diverse data distributions, and the design is inherently compatible with multimodal extensions like LiDAR or thermal imaging. While the multi-branch architecture supports scalability to high-resolution data, memory consumption remains a trade-off that may be mitigated by patch-wise inference. Lightweight backbones, pruning, and knowledge distillation further offer avenues to reduce computational cost, enabling large-scale deployment. Future work should explore integrating multimodal inputs, such as natural language descriptions or user-provided sketches, to guide restoration and improve semantic coherence and visual realism. Recent advances combining Transformer-based language models [49,50] with generative frameworks indicate promising directions. Additionally, ongoing efforts in model compression and optimization remain essential to enable real-time deployment on edge devices. Emerging techniques incorporating reinforcement learning and interactive human–machine collaboration may further personalize and enhance image restoration services. Moreover, although the Agent Attention mechanism has been validated primarily on urban datasets in this study, its formulation is domain-agnostic and, in principle, applicable to indoor and natural environments. Future work will extend evaluation to these scenarios (e.g., NYUv2, ScanNet, ADE20K) to rigorously assess generalization capability across diverse domains. In addition, the framework could be extended by directly benchmarking against diffusion-based and advanced GAN-based inpainting methods. While diffusion approaches often achieve strong visual realism, they are computationally expensive, raising challenges for real-time deployment. A systematic comparison of realism versus computational efficiency will provide deeper insight into practical trade-offs.

In conclusion, this research offers a robust multi-scale interaction framework that advances the quality and interpretability of image inpainting. Yet, realizing its full potential in practical settings necessitates overcoming challenges related to computational efficiency,

ethical safeguards, and multimodal integration. Addressing these issues is critical to translating theoretical progress into real-world impact.

Author Contributions: Conceptualization, Y.L.; Methodology, Y.L.; Software, Y.L.; Validation, A.P.; Data curation, K.C.; Writing—original draft, Y.L.; Writing—review & editing, K.C.; Supervision, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
2. Chan, T.F.; Shen, J. Nontexture inpainting by curvature-driven diffusions. *J. Vis. Commun. Image Represent.* **2001**, *12*, 436–449. [CrossRef]
3. Bertalmio, M.; Vese, L.; Sapiro, G.; Osher, S. Simultaneous structure and texture image inpainting. *IEEE Trans. Image Process.* **2003**, *12*, 882–889. [CrossRef] [PubMed]
4. Pei, Z.; Jin, M.; Zhang, Y.; Ma, M.; Yang, Y.-H. All-in-focus synthetic aperture imaging using generative adversarial network-based semantic inpainting. *Pattern Recognit.* **2021**, *111*, 107669. [CrossRef]
5. Zou, Q.; Cao, Y.; Li, Q.; Mao, Q.; Wang, S. Automatic inpainting by removing fence-like structures in RGBD images. *Mach. Vis. Appl.* **2014**, *25*, 1841–1858. [CrossRef]
6. Han, X.; Zhang, Z.; Du, D.; Yang, M.; Yu, J.; Pan, P.; Yang, X.; Liu, L.; Xiong, Z.; Cui, S. Deep reinforcement learning of volume-guided progressive view inpainting for 3D point scene completion from a single depth image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 234–243.
7. Xiang, H.; Zou, Q.; Nawaz, M.A.; Huang, X.; Zhang, F.; Yu, H. Deep learning for image inpainting: A survey. *Pattern Recognit.* **2022**, *134*, 109046. [CrossRef]
8. Xie, J.; Xu, L.; Chen, E. Image denoising and inpainting with deep neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, USA, 3–8 December 2012; Volume 25.
9. Qu, Z.; Garfinkel, A.; Weiss, J.N.; Nivala, M. Multi-scale modeling in biology: How to bridge the gaps between scalesscales? *Prog. Biophys. Mol. Biol.* **2011**, *107*, 21–31. [CrossRef] [PubMed]
10. Fawzi, A.; Samulowitz, H.; Turaga, D.; Frossard, P. Image inpainting through neural networks hallucinations. In Proceedings of the Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Bordeaux, France, 11–12 July 2016; pp. 1–5.
11. Zhang, J.; He, L.; Karkee, M.; Zhang, Q.; Zhang, X.; Gao, Z. Branch detection for apple trees trained in fruiting wall architecture using depth features and Regions-Convolutional Neural Network (R-CNN). *Comput. Electron. Agric.* **2018**, *155*, 386–393. [CrossRef]
12. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Generative image inpainting with contextual attention. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 5505–5514.
13. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [CrossRef]
14. Bertalmio, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
15. Criminisi, A.; Pérez, P.; Toyama, K. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process* **2004**, *13*, 1200–1212. [CrossRef] [PubMed]
16. Beya, O.; Hittawe, M.; Sidibé, D.; Mériaudeau, F. Automatic detection and tracking of animal sperm cells in microscopy images. In Proceedings of the 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Bangkok, Thailand, 23–27 November 2015; pp. 155–159.
17. Hays, J.; Efros, A.A. Scene completion using millions of photographs. *ACM Trans. Graph.* **2007**, *26*, 4. [CrossRef]
18. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 107. [CrossRef]
19. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.

20. Park, T.; Liu, M.-Y.; Wang, T.-C.; Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 2337–2346.
21. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212. [CrossRef]
22. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
23. Burt, P.J.; Adelson, E.H. The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **1983**, *31*, 532–540. [CrossRef]
24. Ronneberger, Ö.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
25. Hittawe, M.M.; Sidibé, D.; Mériaudeau, F. Bag of words representation and SVM classifier for timber knots detection on color images. In Proceedings of the 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Tokyo, Japan, 18–22 May 2015; pp. 287–290.
26. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
27. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 5693–5703.
28. Köhler, R.; Schuler, C.; Schölkopf, B.; Harmeling, S. Mask-specific inpainting with deep neural networks. In Proceedings of the German Conference on Pattern Recognition, Düsseldorf, Germany, 7–10 September 2014; pp. 523–534.
29. Yu, Y.; Zhan, F.; Wu, R.; Pan, J.; Cui, K.; Lu, S.; Ma, F.; Xie, X.; Miao, C. Diverse image inpainting with bidirectional and autoregressive transformers. In Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), Virtual Event, 20–24 October 2021.
30. Paris StreetView Dataset. CSA-Inpainting. v3. Available online: https://github.com/KumapowerLIU/ (accessed on 8 September 2025).
31. Liu, H.; Jiang, B.; Xiao, J.; Yang, C. Coherent Semantic Attention for Image Inpainting. *arXiv* **2019**, arXiv:1905.12384v3. [CrossRef]
32. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.-C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Cambridge, UK, 19–22 September 2018; pp. 85–100.
33. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.; Ebrahimi, M. EdgeConnect: Structure guided image inpainting using edge prediction. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3265–3274.
34. Pan, X.; Zhan, X.; Dai, B.; Lin, D.; Loy, C.C.; Luo, P. Exploiting deep generative prior for versatile image restoration and manipulation. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 262–277.
35. Guo, Q.; Li, X.; Juefei-Xu, F.; Yu, H.; Liu, Y.; Wang, S. JPGNet: Joint predictive filtering and generative network for image inpainting. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 386–394.
36. Zhang, H.; Hu, Z.; Luo, C.; Zuo, W.; Wang, M. Semantic image inpainting with progressive generative networks. In Proceedings of the 26th ACM International Conference on Multimedia (ACM MM), Seoul, Republic of Korea, 22–26 October 2018; pp. 1939–1947.
37. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. In *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*; Wiley: Hoboken, NJ, USA, 2019; Volume 9, p. e1301.
38. Ramzan, M.; Abid, A.; Bilal, M.; Aamir, K.M.; Memonand, S.A.; Chung, T.-S. Effectiveness of pre-trained CNN networks for detecting abnormal activities in online exams. *IEEE Access* **2024**, *12*, 21503–21519. [CrossRef]
39. Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; Huang, T.S. Free-form image inpainting with gated convolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4470–4479.
40. Avazov, K.; Jamil, M.K.; Muminov, B.; Abdusalomov, A.B.; Cho, Y.-I. Fire detection and notification method in ship areas using deep learning and computer vision approaches. *Sensors* **2023**, *23*, 7078. [CrossRef] [PubMed]
41. Yeh, R.A.; Chen, C.; Yian, T.; Lim, A.G.; Schwing, M.; Hasegawa-Johnson, M.N. Do Semantic image inpainting with deep generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6882–6890.
42. Wang, N.; Ma, S.; Li, J.; Zhang, Y.; Zhang, L. Multistage attention network for image inpainting. *Pattern Recognit.* **2020**, *106*, 107448. [CrossRef]

43. Sagong, M.-C.; Shin, Y.-G.; Kim, S.-W.; Park, S.; Ko, S.-J. PEPSI: Fast image inpainting with parallel decoding network. In Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 11360–11368.
44. Li, J.; Wang, N.; Zhang, L.; Du, B.; Tao, D. Recurrent feature reasoning for image inpainting. In Proceedings of the IEEE/CVF IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7757–7765.
45. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008.
46. Ren, J.S.; Xu, L.; Yan, Q.; Sun, W. Shepard convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; Volume 28, pp. 907–915.
47. Dapogny, A.; Cord, M.; Pérez, P. The missing data encoder: Cross-channel image completion with hide-and-seek adversarial network. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 10688–10695. [CrossRef]
48. Xie, C.; Liu, S.; Li, C.; Cheng, M.-M.; Zuo, W.; Liu, X.; Wen, S.; Ding, E. Image inpainting with learnable bidirectional attention maps. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8857–8866.
49. Ma, Y.; Liu, X.; Bai, S.; Wang, L.; He, D.; Liu, A. Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI), Macau, China, 10–16 August 2019; pp. 3123–3129.
50. Cai, W.; Wei, Z. PiGAN: Generative adversarial networks for pluralistic image inpainting. *IEEE Access* **2020**, *8*, 48451–48463. [CrossRef]