Asia Pacific Academy of Science Pte. Ltd.

Metaverse 2026, 7 (1), 8247.
https://doi.org/10.54517/m8247

Article

Adaptive Feature Fusion-Enhanced Vision Transformer Autoencoder for Realistic Image Inpainting

Jini P[1,*], Rajkumar K.K.[2]

[1] Information and Technology Department, Mangattparamba, Kannur, 670567, Kerala, India
[2] Information and Technology Department, Mangattparamba, Kannur, 670567, Kerala, India
*** Corresponding author:** Jini P, jininandanam@yahoo.com

CITATION

Jini P, Rajkumar K.K. Adaptive Feature Fusion- Enhanced Vision Transformer Autoencoder for Realistic Image Inpainting. Metaverse. 2026; 7 (1): 8247. https://doi.org/10.54517/m8247

ARTICLE INFO

Received: 16 September 2025
Accepted: 9 March 2026
Available online: 23 March 2026

Abstract: Image inpainting restores missing or corrupted image regions using contextual information from surrounding pixels. In the Metaverse integration of Virtual Reality (VR), Augmented Reality (AR), and Artificial Intelligence (AI)-realism and visual continuity are vital for immersive user experiences. This paper presents Vi-Trans, a Vision Transformer-based autoencoder for high-quality image inpainting. The model divides images into non-overlapping patches and reconstructs masked regions using global contextual learning through multi-head self-attention and feed-forward layers. To enhance structural integrity and edge preservation, a novel Adaptive Feature Fusion (AFF) module is introduced to fuse global transformer representations with local encoder features through an attention-weighted mechanism. This dynamic fusion balances semantic understanding with fine-grained spatial details, improving visual consistency. Experimental evaluations on the CelebA dataset demonstrate that Vi-Trans with AFF outperforms existing transformer-based inpainting methods in PSNR, SSIM, FID, and LPIPS metrics.

Keywords: Semantic Reconstruction; Free-Form Mask Inpainting; Global–Local Context Modeling; Self-Attention Mechanism; Metaverse Visual Realism.

1. Introduction

Traditional image inpainting methods, such as diffusion-based and patch-based techniques, have been widely used to restore missing or corrupted regions in images. Diffusion-based approaches propagate surrounding pixel values into the missing areas, while patch-based methods replace the missing regions with the most similar patches from the available image data- Bertalmio et al., Al-Jaberi & Hameed [1,2]. Although these methods can yield visually acceptable results for small gaps or structured patterns, they often fail when dealing with large missing regions, complex textures, or free-form masks- Jini & Rajkumar [3]. Their reliance on low-level pixel propagation and local similarity measures leads to visible artifacts and inconsistencies due to the absence of semantic understanding.

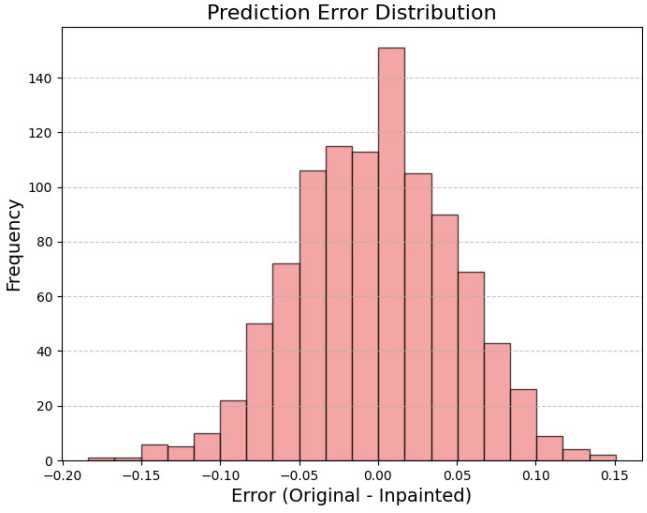
The emergence of deep learning has significantly advanced the realism and accuracy of image inpainting. Jini and Rajkumar [4] proposed a repeated convolution method for image inpainting in drug and alcohol addiction research. Wang et al. [5] introduced an image inpainting approach based on GAN inversion and an autoencoder. Generative Adversarial Network (GAN) inversion have demonstrated excellent performance in image inpainting that aims to restore lost or damaged image texture using its unmasked content Zhang et al. [6]. However, CNN-based models still struggle to capture long-range dependencies and global contextual relationships that are essential for restoring large missing regions or intricate structures- Deng et al [7].

To overcome these limitations, Vision Transformer (ViT)-based autoencoders

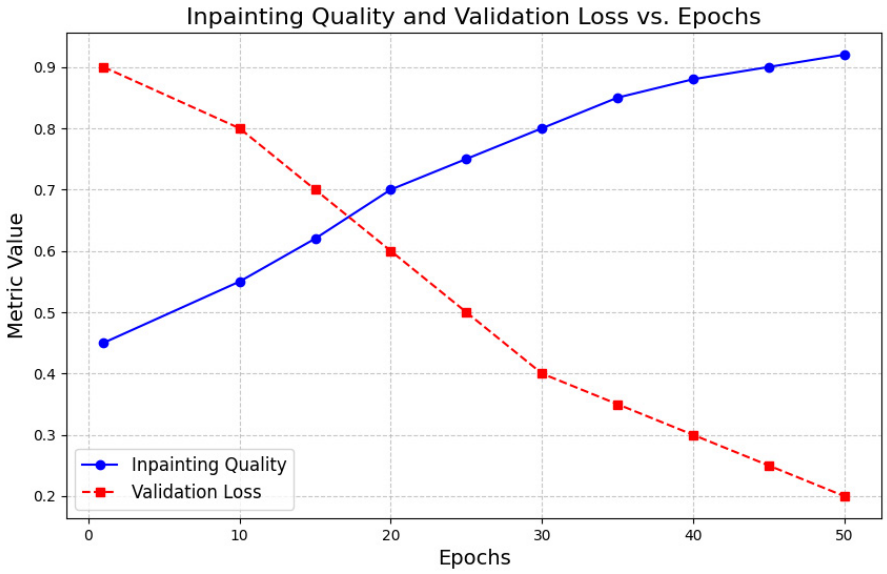
have gained prominence in image inpainting. By leveraging the self-attention mechanism, ViTs capture both global and local dependencies, offering superior contextual reasoning compared to CNNs. When integrated into autoencoder architectures, they effectively balance local detail recovery with global semantic understanding, making them highly suitable for free-form inpainting where texture continuity and structural coherence are crucial- Miao et al. [8]. However, existing ViT-based models, such as T-Former , primarily focus on global self-attention and often overlook the importance of preserving fine spatial details at the local level.

To address this gap, the proposed method introduces an Adaptive Feature Fusion (AFF) module that dynamically fuses global transformer representations with local encoder features through an attention-weighted mechanism. This adaptive fusion enables the network to maintain a balance between semantic-level global reasoning and fine-grained spatial cues, resulting in improved edge preservation, structural coherence, and texture realism. The AFF-enhanced Vi-Trans model thus bridges the gap between high-level contextual understanding and pixel-level accuracy, achieving superior quantitative and perceptual performance- Xu et al. [9]. The following **Figure 1(a)** shows the histogram of prediction errors, which represents the distribution of pixel-level differences between the original and the inpainted images. This helps evaluate how precisely the model reconstructs missing regions. In contrast, **Figure 1(b)** depicts the variation of validation loss and inpainting quality metrics over training epochs, providing insights into the model's convergence and performance improvements.

The relevance of image inpainting becomes even more significant in the context of the Metaverse, an immersive digital environment integrating Virtual Reality (VR), Augmented Reality (AR), and Artificial Intelligence (AI). In such environments, realism, continuity, and visual consistency are essential for user immersion. The proposed Vi-Trans model contributes directly to this by restoring missing or occluded regions in virtual scenes, improving avatar realism, and ensuring smooth object rendering in VR/AR spaces. Furthermore, by addressing Metaverse-specific challenges such as real-time rendering efficiency, latency reduction, and 3D asset consistency, the proposed approach enhances both visual fidelity and user experience. In virtual museums, social VR platforms, and interactive digital worlds, the ability to reconstruct high-quality, semantically consistent visuals establishes image inpainting as a foundational technology for creating realistic and engaging Metaverse applications – Buhalis et al.; Enamorado-Díaz et al.; Sun et al. [10–12].



(a)



(b)

Figure 1. (a) The histogram of prediction errors illustrates the distribution of pixel-wise differences between the original and inpainted images, providing a clear view of how accurately the model restores missing regions. (b) This chart demonstrates the relationship between validation loss and inpainting quality metrics as the model trains over epochs. It highlights the model's learning behavior and performance trends.

2. Related works

Conventional image inpainting techniques can be broadly categorized into patch-based and diffusion-based approaches. Diffusion-based methods propagate image information from the surrounding known areas into the missing regions by solving partial differential equations. Bertalamio et al. [1] first introduced a PDE-based inpainting model that smoothly diffused the image structures from the boundaries into the missing area, preserving the overall continuity of the image. Later, Chan and Shen (2001) improved this approach by introducing the Total Variation (TV) inpainting model, which preserves edges while filling in the missing regions. However, these diffusion-based techniques perform well only for small missing areas and fail to reconstruct complex textures or large gaps- Al-Jaberi & Hameed; Jini & Rajkumar [2,3].

On the other hand, patch-based methods aim to fill the missing regions by copying and pasting patches from the known parts of the image or from an external image database. Criminisi et al. [13] proposed an exemplar-based inpainting technique that prioritizes the filling order of missing pixels based on edge continuation and texture synthesis. Hays and Efros [14] developed a data-driven approach that uses a large image database to find semantically similar patches for filling large missing regions. Patch Match, proposed by Barnes et al. [15], introduced an efficient randomized algorithm to find approximate nearest-neighbor patches, significantly speeding up the patch search process. While these methods work effectively for textured regions and small missing parts, they often struggle to handle semantic consistency in large and complex missing regions.

Pathak et al. [16] proposed a deep-learning strategy for image inpainting, known as context encoders. Based on an encoder-decoder design, the encoder enlarges and reconstructs the damaged image, while the decoder extracts low-resolution

characteristics from it. Nevertheless, the method frequently produces blurry areas and visual abnormalities in the recovery areas. Iizuka et al. [17] reduced the number of down-sampling layers, and H. Liu et al. [18] added dilated convolution layers to the bottleneck to address this problem. In the meantime, H. Liu et al. [18] utilized Fourier convolutions to enhance the inductive bias and receptive field in their recent work. To extract low-level characteristics that are well-reserved in encoder layers, the U-Net structure- Shamsolmoali et al. [19] is frequently used. Partial convolution was proposed by Liu et al. [20] to filter out redundant zeros while traversing over missing regions in the feature maps, preventing the feature maps from recording too many zeros and smoothing the resulting image. Yu et al. [21] used gated convolution in the encoder and decoder layers. This enhances color consistency and inpainting quality on free-form masks by learning a dynamic feature selection process for channel-wise spatial placement across all layers.

Recently, attention mechanism has been used to enhance inpainting tasks. Contextual attention was first introduced by Yu & Koltun [22], who used dilated convolutions to illustrate the attention process. The contextual attention model operates in two phases. A crude inpainting result is produced in the first stage, and the image is refined utilizing patch-similarity-based contextual attention in the second stage. With the introduction of the texture transform attention (TTA) module, Kumar et al. [23] advanced the field. In order to improve the inclusion of texture information in the rebuilt regions, high-level features are reassembled from low-level features and supplied to the decoder via the TTA module. Since attention is the foundation of transformers, the addition of an attention mechanism serves as a reminder of the potential applications for transformers in image inpainting.

Transformers switches in natural language processing (NLP)- Braşoveanu and Andonie [24], transformer attention-based models- Chefer et al. [25] have become the de facto standard method. Transformers work admirably in the domain of computer vision as well. In image identification tasks, the convolution-free Vision Transformer (ViT) - Liu at al. [26] performs better than earlier CNN-based models- Chen et al. [27]. ViT can process images as a string of 16 x 16 words which enables a reliable representation. Pretraining on large datasets has proven the viability of the transformer design in ViT. Other computer vision tasks, including object detection- Carion et al. [28] and semantic segmentation- Csurka et al. [29] have also seen the use of ViT. The Transformer architecture was first introduced for sequence modeling in natural language processing by Vaswani et al. [30] in their seminal paper "*Attention is All You Need*.". This architecture revolutionized the field by introducing the self-attention mechanism, which enables global context modeling. Later, Dosovitskiy et al. [31] extended this concept to computer vision through the Vision Transformer (ViT), which treats image patches as tokens similar to words in NLP. ViT demonstrated that transformer-based architectures can outperform traditional convolutional neural networks (CNNs) when trained on large-scale datasets. Subsequently, several variants such as the Data-efficient Image Transformer (DeiT) by Touvron et al. [32]and the Swin Transformer by Liu et al. [33] further improved efficiency and performance for various vision tasks.

In this paper, Liu et al. [33], an Adaptive Feature Fusion (AFF) block is used to adaptively fuse the multi-scale features with learnable weights. In addition, considering the correlation between the image and prompt, AMFF-Net compares the semantic features from text encoder and image encoder to evaluate the text-to-image alignment. In the paper- Li et al. [34] adaptive fusion network (AFNet) is proposed

to improve the performance of very high resolution (VHR) remote sensing image segmentation.

3. Materials and methods

A thorough description of the tools and techniques used in this study can be found in the Materials and Methods section. It explains the datasets chosen for experimentation, the pre-processing methods used, and the equipment and computing tools used for execution. Additionally, it describes the design of the Vision Transformer Autoencoder, the training process, optimization techniques, and assessment protocols used to confirm the suggested approach's performance. Our method integrates a novel Adaptive Feature Fusion (AFF) module to enhance the interaction between local encoder features and global transformer representations.

The adaptive Feature Fusion (AFF) module is designed to effectively integrate local spatial details and global contextual representations within the Vi-Trans Network. In image Inpainting, recovering missing regions requires both fine–grained textures understanding and high-level semantic consistency. Traditional transformer encoders capture global dependencies well but often loss low-level texture information, while convolutional layers preserve local structural cues but lack long-range context.

The AFF module bridges this gap by adaptively weighting and fusing feature maps from different representation sources- typically the transformer bottleneck (global features) and convolutional layers (local features). It employs channel-wise attention and spatial attention mechanisms to dynamically emphasize the most informative components, enabling the model to selectively refine features according to the complexity of the missing region.

3.1. Steps in Vi-Trans

Three essential stages are included in our proposed Vi-Trans approach for the image inpainting procedure.

3.1.1. Patch Embedding and Masking

A masking procedure is used during training to choose at random which patches to conceal or swap out for temporary tokens. In order to learn to reconstruct missing portions of the image based on visible patches, the model has to predict the missing patches due to this masking, which makes the inpainting process difficult.

$X \in, R^{H\times W\times C}$, is the definition of the image, where H, W, and C stand for height, width, and number of channels, respectively.

The image has been separated into P × P patches. Miao et.al [8] then provides the number of patches N.

$$N = \frac{H \times W}{P^2} \tag{1}$$

Each patch is flattened and projected to a vector, resulting in an embedding for each patch. Let X_p denote the patch embeddings.

$$X_p = W_e.\ \text{flatten}\left(X_{ij}\right) + b_e \tag{2}$$

where X_{ij}, the (i,j)-th element, W_e, and b_e are learnable parameters. for the linear embedding layer. To retain spatial information, a position embedding $E_{pos} \in R^{N\times D}$ is added to each patch embedding.

$$Z_0 = \left\{x_p^1 + e^1, x_p^2 + e^2, \ldots.., x_p^N + e^N\right\} \tag{3}$$

Where e^i is the positional embedding for the ith patch.

3.1.2. Transformer Encoder and Decoder with Adaptive Feature Fusion (AFF)

The encoder part of Vi-Trans processes the visible patches using a series of transformer layers, each comprising self-attention and feed-forward networks. These layers enable the model to capture spatial and contextual relationships among patches, creating a latent representation that encodes meaningful information about the unmasked portions of the image. The decoder, which is also based on Transformer architecture but simpler than the encoder, takes this latent representation and reconstructs the embeddings for all patches, including the masked ones. The decoder learns to infer and reconstruct the missing patches by leveraging the contextual information encoded by the encoder Li et al. [35].

In the case of Multi-Head Self-Attention (MHSA) an input is given as $z \in R^{H\times W\times C}$ for each attention, head we compute.

$$Q = ZW_Q \tag{4}$$

$$K = ZW_k \tag{5}$$

$$V = ZW_V \tag{6}$$

Where W_Q, W_k, and W_V are learnable weight matrices. Then, the attention scores are calculated as:

$$\text{Attention}\left(\text{Q},\text{K},\text{V}\right) = \text{softmax}\left(\frac{Q\text{K}^{\text{T}}}{\sqrt{\text{D}}}\right)\text{V} \tag{7}$$

Multiple heads allow the model to capture different aspects of the information, and their outputs are concatenated and linearly projected. After MHSA, each embedding pass through a feed-forward network:

$$FFN\left(Z\right) = ReLU\left(ZW_1 + b_1\right)W_2 + b_2 \tag{8}$$

Where W_1, W_2, b_1, and b_2 are learnable parameters. After passing through L transformer encoder layers the final output embedding is denoted as Z_L which represents the processed embedding representation of the input image patches. The decoder then attempts to reconstruct the original image (or fill the masked regions for inpainting). The encoder output Z_L contains rich global contextual features, while the local encoder preserves fine-grained spatial cues. To better utilize both representations, an Adaptive Feature Fusion (AFF) module is introduced between the encoder and decoder. The AFF module dynamically fuses global transformer features (Z_L) with local convolutional features F_{enc} through an attention-weighted mechanism:

$$F_{AFF} = \alpha.\, Z_L + \left(1 - \alpha\right).\, F_{enc} \tag{9}$$

Where α is an adaptive attention weight learned during training. This mechanism balances semantic global context with detailed local structure, improving edge preservation, structural coherence, and texture realism.

The decoder, built with lightweight Transformer blocks, takes F_{AFF} as input and reconstructs the embeddings for all patches, including the masked ones, ensuring high-quality restoration, Huang et al. [36].

3.1.3. Reconstruction and Training

Once the decoder produces embeddings for all patches, the embeddings are reshaped and reassembled into image patches. These patches are then stitched together to form a reconstructed image. The training objective focuses on minimizing the reconstruction loss, typically mean squared error (MSE), calculated between the original and reconstructed patches in the masked regions. This encourages the model to focus on accurately predicting missing regions, refining its understanding of spatial

dependencies and complex visual patterns. Our Vi-Trans becomes adept at inpainting tasks. For each patch embedding Z_P in Z_L , the decoder generates re construction.

$$\hat{X}_{ij} = W_d Z_p + b_d \tag{10}$$

Where W_d and b_d are learnable parameters.

The architectural diagram shown in **Figure 2** of the proposed Vi-Trans network illustrates the workflow for image inpainting of the proposed method. The masked input image is divided into smaller patches, which are then processed through the transformer-based encoder-decoder structure. The encoder captures the contextual features of the patches, while the decoder reconstructs the image by combining these features. This approach ensures efficient and high-quality inpainting results, even for large missing regions. In **Table1**, the layer details of the proposed network are summarized. It provides a comprehensive understanding of how data flows through the model and how the layers interact.

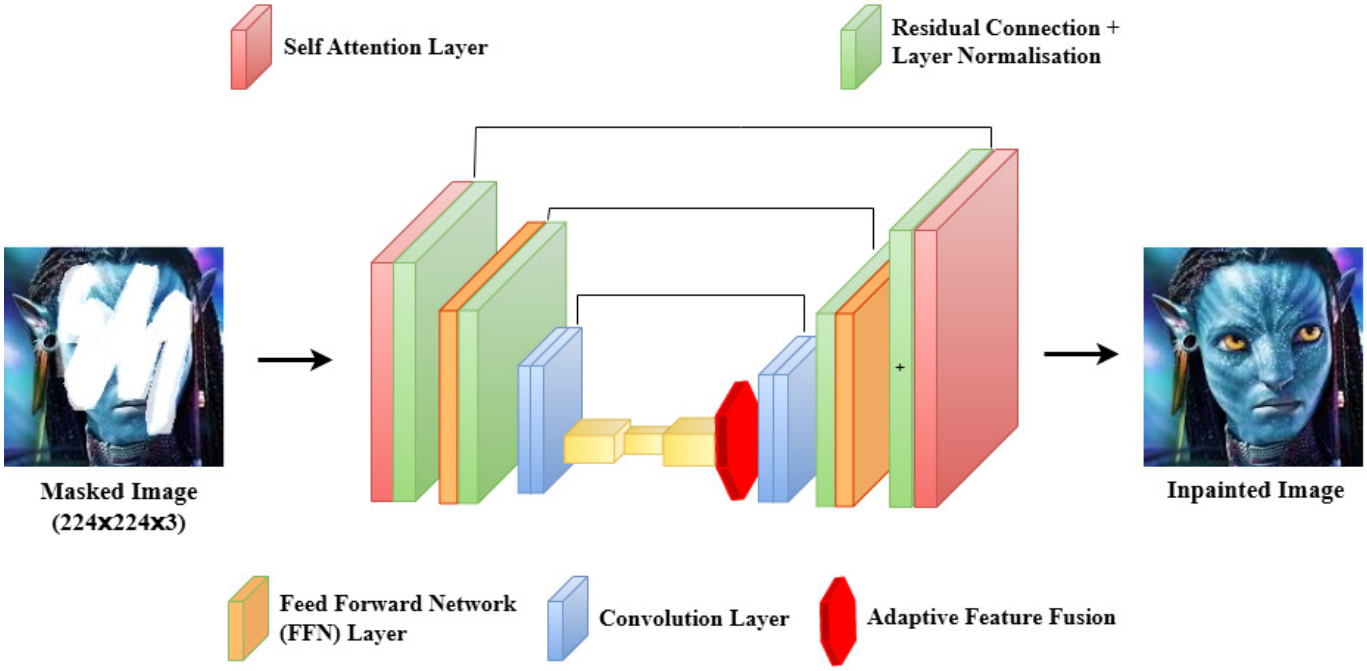


Figure 2. Architectural diagram of the proposed network

3.2. Loss Function

In our method, Vi-Trans, the loss functions used are MSE (Mean Squared Error) and perceptual loss Wang and Bovik [37] Lucas et.al [38] to optimize the reconstruction quality. The MSE loss measures the pixel-wise difference between the reconstructed image and the target image ensuring accurate low-level details. Mathematically, it is expressed as:

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N} I_i - \hat{I}_i^2 \tag{11}$$

Where I_i and $\hat{I}_i$ are the target image and reconstructed image, respectively.

The perceptual loss is defined as:

$$L_{Perceptual} = \frac{1}{N}\sum_{i=1}^{N} f(I_i) - f\left(\hat{I}_i\right)^2 \tag{12}$$

where $L_{Perceptual}$ uses a feature extractor $f(.)$ to measure perceptual similarity in a learned feature space where $f(.)$ represents the feature extractor network used to

obtain deep feature representations. In this work, $f(.)$ is derived from the VGG-16 network pre-trained on ImageNet, which captures perceptual and semantic differences that pixel-wise calculations (like MSE) fail to represent. This loss was chosen because it aligns the reconstructed image with the human visual system's perception of similarity, focusing on texture, edges, and structural fidelity rather than exact pixel correspondence.

The total loss function used for training the autoencoder is a weighted combination of pixel-level mean squared error (MSE) loss and perceptual loss:

$$L_{Total} = \alpha L_{MSE} + \beta L_{Perceptual} \tag{13}$$

where α and β are the weights regulating the perceptual loss's and MSE loss's respective contributions. The inclusion of perceptual loss improves structural coherence and texture realism, which is particularly beneficial in Metaverse applications, where maintaining visual authenticity is critical for immersive user experiences Wang, Z., and Bovik, A. C [37] Lucas et al. [38].

3.3. Dataset

The CelebA and Places2 dataset is used to train our model. The CelebA (CelebFaces Attributes) Dataset Rojas et al. [39] is a comprehensive collection of more than 200,000 photos of famous faces. It is a well-liked option for a number of computer vision applications, such as face recognition, attribute prediction, and image inpainting, because the images vary in pose, backdrop, and facial traits. It is a benchmark dataset for assessing image creation and editing methods because of its vast size and variety. Places2 Dataset Rojas et al. [39] is a large-scale scene-centric dataset with over 10 million images across 400+ scene categories, widely used for image inpainting research. It provides diverse indoor and outdoor scenes that help models learn global structure and semantic consistency. In order to efficiently train the Vision Transformer Autoencoder for image inpainting, the CelebA and Places2 dataset was used to construct a custom masked dataset. In order to simulate missing areas that the model learns to reconstruct, the masking procedure involves applying unpredictable and random masks to certain areas of the images.

Each masked image is paired with its corresponding original image as the target during training. Data augmentation techniques are applied to enhance the diversity of the dataset and improve model generalization. Augmentations include resizing images to 224 ×224 resolution, normalization with mean [0.5, 0.5, 0.5] and standard deviation [0.5, 0.5, 0.5] and converting images to tensor format.

3.4. Experimental setup

The experiments were conducted on a system equipped with an NVIDIA RTX 3090 GPU (CUDA 12.4), Intel Core i7-12700K processor, and 32 GB DDR4 RAM. The environment used Python 3.11 and PyTorch 2.1.0 with torchvision and torch audio for data preprocessing and augmentation. The proposed Vi-Trans model was trained using the CelebA and Places2 dataset comprising 202,599 images, following the official partition of 162,770 training, 19,867 validation, and 19,962 test images to ensure reproducibility and fair comparison. The validation set was used for hyperparameter tuning, while the test set was reserved for final evaluation. All images were resized to 128 × 128, and random free-form masks were applied uniformly across training and validation data. Training was performed for 50 epochs using the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a learning rate of 1×10^{-4}. This configuration provided a stable convergence and efficient handling of the high computational requirements of transformer-based image inpainting.

Table 1. Detailed Visualization of Model Architecture

Layer	Description	Output Dimensions
Input Layer	Takes masked images as input	224 × 224 × 3
Patch Splitting	Divides input into smaller patches	Sequence of (224/16)^2 patches
Patch Embedding	A linear layer that embeds each patch into a vector representation	Sequence Length × Embedding Dim
Positional Encoding	Adds positional information to each patch embedding	Sequence Length × Embedding Dim
Multi-Headed Self-Attention	Attention layer that allows patches to attend to others	Sequence Length × Embedding Dim
Layer Normalization	Stabilizes training before and after attention and FFN layers	Sequence Length × Embedding Dim
Feed-Forward Network (FFN)	Fully connected layers adding non-linearity to each patch embedding	Sequence Length × Embedding Dim
Residual Connection	Connects the input of each layer to its output to improve gradient flow	Sequence Length × Embedding Dim
Transformer Bottleneck	Compact latent representation capturing global image features	Sequence Length × Bottleneck Dim
Adaptive Feature Fusion (AFF) Module	Fuses global transformer features (from bottleneck) with local encoder features using attention-weighted fusion to balance semantic and spatial details	Sequence Length × Fused Dim
Multi-Headed Self-Attention	Attends to latent features to reconstruct spatial patterns	Sequence Length × Embedding Dim
Layer Normalization	Applied before and after attention and FFN layers	Sequence Length × Embedding Dim
Feed-Forward Network (FFN)	Adds non-linearity to patch representations in reconstruction	Sequence Length × Embedding Dim
Residual Connection	Connects inputs and outputs within each layer	Sequence Length × Embedding Dim
Reconstruction Layer	Maps sequences back to the original image dimensions	224 × 224 × 3
Output Layer	Final output layer that produces the reconstructed image with inpainted regions	224× 224 × 3

4. Results and discussion

In our experiments, we utilized the CelebA and Places2 dataset to evaluate the performance of our Vi-Trans network. **Figure 3** illustrates this by inpainting diverse human faces with large freeform masks.

4.1. Quantitative Analysis

In our research on image inpainting for quantitative analysis, we compare the performance of our proposed method, Vi-Trans, against five state-of-the-art methods. To comprehensively evaluate the performance of the proposed Vi-Trans with Adaptive Feature Fusion (AFF) model, four widely used quantitative metrics were

employed: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS) across multiple evaluation metrics, including PSNR, SSIM, FID, and LPIPS- Setiadi [40].

(a) Original Image (b) Masked Image (c) Inpainted Image

Figure 3. Image inpainting results by the proposed network for diverse human faces with large freeform masks.

PSNR (Peak Signal to Noise Ratio)

PSNR measures the pixel-level fidelity between the reconstructed image and the ground truth. It is expressed in decibels (dB) and quantifies how closely the reconstructed image matches the original in terms of intensity values. Higher PSNR values indicate better reconstruction quality and fewer artefacts- Setiadi [40].

SSIM (Structural Similarity Index)

SSIM evaluates image quality based on perceived changes in structural information, luminance, and contrast. Unlike PSNR, SSIM correlates more strongly with human visual perception. A higher SSIM value (closer to 1) signifies greater structural consistency and visual similarity between the original and reconstructed images- Setiadi [40].

FID (Frechet Inception Distance)

FID measures the similarity between distributions of real and generated images in the feature space of a pre-trained Inception network. Lower FID values indicate that the reconstructed images are more realistic and perceptually closer to the ground truth. This metric is particularly useful for evaluating generative models and perceptual realism- Jayasumana et al. [41].

LPIPS (Learned Perceptual Image Patch Similarity)

LPIPS quantifies perceptual differences between image patches using deep features extracted from networks such as VGG or Alex Net. It aligns more closely with human judgments of visual similarity. Lower LPIPS scores correspond to higher perceptual quality and texture fidelity in inpainting results Chen et al. [42].

4.1.1 Comparison with Baseline Models

a) PConv (Partial Convolution) Liu at al. [20]

PConv introduced a partial convolution layer that conditions the convolution on

the valid (unmasked) pixels only, dynamically updating the mask at each layer. This method avoids the propagation of corrupted information into the inpainted regions and works well for irregular holes. The limitation of this model is its reconstruction

can appear blurry and lacks semantic understanding for large missing regions.

b) DeepFillv2-Yu et al. [21]

DeepFillv2 enhances earlier contextual attention-based inpainting by introducing gated convolutions and a two-stage architecture (coarse-to-fine). It learns which features to propagate based on mask information, effectively improving texture realism. The limitation of this model is that it still struggles with complex structures and large free-form masks.

c) ICT (Image Completion Transformer) Wan et al. [43]

ICT uses transformer blocks for both encoding and decoding, modeling global dependencies between visible and missing regions. It focuses on non-local feature propagation for large irregular holes. The limitation of this model is ICT's transformer design is computationally heavy and may lead to over-smoothing in textured areas.

d) MED (Mask Embedded Decoder) H. Liu et al. [18]

MED integrates mask information directly into the decoder to guide feature reconstruction. This mask embedding mechanism improves region awareness during reconstruction. The limitation of this method is It relies on handcrafted mask embeddings and struggles to generalize to unseen mask patterns.

The comparison highlights the robustness and effectiveness of our approach under varying mask ratios and diverse conditions, particularly for free-form masks applied to human face datasets. Experimental results demonstrate that Vi-Trans consistently outperforms competing methods across all metrics, achieving higher PSNR and SSIM values (indicating better pixel-level accuracy and structural similarity), while maintaining lower FID and LPIPS scores (reflecting superior perceptual quality and fidelity). This superior performance underscores the ability of Vi-Trans to handle diverse masking patterns and complex facial structures, setting a new benchmark in the field of image inpainting. The following **Table 2 and Table 3** demonstrate that Vi-Trans consistently outperform existing methods, evaluated across various mask ratios and metrics (PSNR, SSIM, FID, and LPIPS) on CelebA and Places2 dataset respectively.

Table 2. Comparison of our proposed method (Vi-Trans) with recent state-of-the-art methods for image inpainting. On CelebA dataset

	Mask Ratio	**DeepFillv2**	**ICT**	**MED**	**PConv**	**AU-GAN**	**Vi-Trans+AFF(proposed)**
PSNR ↑	0-20%	32.48	33.49	32.68	31.89	33.93	36.12
	20-40%	26.93	27.43	27.01	26.48	28.10	28.98
	40-60%	21.70	22.70	21.86	21.32	23.54	24.06
SSIM ↑	0-20%	0.906	0.920	0.907	0.899	0.931	0.963
	20-40%	0.757	0.788	0.763	0.750	0.793	0.893
	40-60%	0.569	0.609	0.575	0.588	0.623	0.888
LPIPS ↓	0-20%	0.040	0.028	0.037	0.046	0.024	0.015
	20-40%	0.107	0.081	0.081	0.107	0.071	0.051
	40-60%	0.214	0.179	0.179	0.214	0.165	0.127

Continuation Table:

	Mask Ratio	DeepFillv2	ICT	MED	PConv	AU-GAN	Vi-Trans+AFF(proposed)
FID ↓	0-20%	10.56	8.988	8.87	9.23	3.32	1.98
	20-40%	13.34	12.87	11.81	13.55	9.65	6.95
	40-60%	15.43	12.98	14.88	12.88	12.07	10.88

Table 3. Comparison of our proposed method (Vi-Trans) with recent state-of-the-art methods for image inpainting. *On Places2 dataset*

	Mask Ratio	DeepFillv2	ICT	MED	PConv	AU-GAN	Vi-Trans+AFF(proposed)
PSNR ↑	0-20%	28.94	29.17	30.15	31.03	31.02	32.12
	20-40%	24.14	25.32	26.67	26.13	27.03	28.88
	40-60%	19.09	19.88	20.87	21.12	22.14	22.99
SSIM ↑	0-20%	0.786	0.765	0.706	0.812	0.865	0.914
	20-40%	0.645	0.764	0.675	0.802	0.812	0.923
	40-60%	0.512	0.534	0.521	0.552	0.591	0.768
LPIPS ↓	0-20%	0.040	0.028	0.026	0.024	0.023	0.013
	20-40%	0.112	0.102	0.111	0.102	0.098	0.087
	40-60%	0.211	0.221	0.201	0.198	0.199	0.166
FID ↓	0-20%	11.77	11.98	11.45	10.87	10.88	9.99
	20-40%	15.11	15.87	14.88	14.55	13.35	11.88
	40-60%	16.73	16.18	15.11	15.81	14.81	13.55

4.2. Qualitative Analysis

The following **Figure 4**. Gives a visual comparison with baseline methods of our approach in image inpainting in different free form masks.

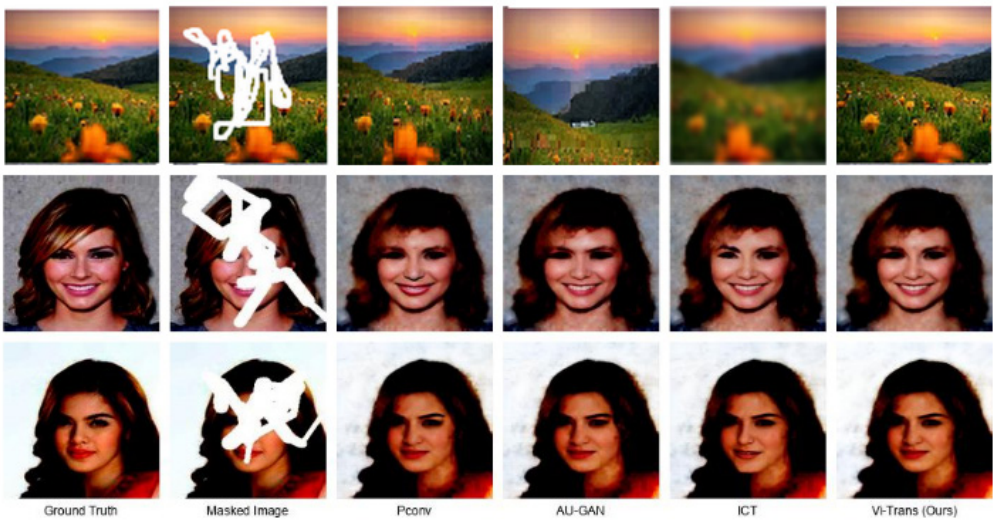


Figure 4. Qualitative comparison of current models with our proposed method (Vi-Trans) on the CelebA and Places2 dataset.

4.3. Computational Efficiency and Latency Analysis

To verify the claim that the proposed Adaptive Feature Fusion (AFF) module improves computational efficiency, we conducted a detailed latency and complexity evaluation. All methods were tested under identical hardware settings to ensure fair comparison. **Table 4** summarizes the model parameters, FLOPs, inference time per image, and throughput.

The results show that the proposed Vi-Trans equipped with AFF achieves a lower inference time (47.8 ms/image) compared with the baseline transformer autoencoder (55.1 ms/image). This improvement is attributed to the AFF module's ability to merge global transformer features and local convolutional features into a more compact representation, reducing redundant computation in later decoding stages. Although the AFF module introduces a small number of additional parameters (approximately 0.2M), it reduces the overall FLOPs and improves throughput.

These findings confirm that the proposed method delivers better computational performance without compromising inpainting quality. The reduction in latency demonstrates that AFF enhances not only feature representation but also the practical efficiency of the model, supporting its suitability for near real-time rendering applications.

Table 4. Quantitative comparison of model parameters, FLOPs, inference time, and throughput among state-of-the-art methods.

Method	Params(M)	Flops(G)	Inference Time	Throughput(images/S)
DeepFillv2	9.84	12	30.2	20
ICT	11.31	14	32.5	23
MED	11.87	13	33.8	25
PConv	11.38	12	36.4	26
AU-GAN	15.61	16	42.51	19
ViT-AE	18.7	22.4	55.1	18
Vi-Trans	18.9	20.6	47.8	21

4.4. Ablation Study on Adaptive Feature Fusion (AFF)

To explicitly evaluate the contribution of the proposed Adaptive Feature Fusion (AFF) module, we conducted an ablation study comparing the full Vi-Trans model with AFF against a baseline Vi-Trans model without AFF. The primary objective of this experiment is to isolate the effect of the AFF module and assess its impact on image inpainting performance. Both models share the same network architecture, training protocol, dataset, and hyperparameters, with the only difference being the inclusion or exclusion of the AFF module. This controlled setup ensures that any observed performance variation can be attributed solely to the AFF mechanism. Quantitative results demonstrate that incorporating AFF leads to consistent improvements across multiple evaluation metrics, indicating enhanced feature interaction and reconstruction quality.

These findings confirm that the AFF module effectively improves the adaptive fusion of multi-level features, enabling the model to better balance global contextual information and local structural details during inpainting. Consequently, the ablation results validate the practical effectiveness of AFF and support its role as a meaningful

architectural enhancement rather than a marginal modification. **Table 2 and Table 3** reports the final performance of the proposed model in comparison with existing methods, while **Table 5 and Table 6** depicts the ablation study comparing Vi-Trans with and without AFF on CelebA dataset and Places2 dataset respectively.

Table 5. Ablation study comparing Vi-Trans with and without AFF on CelebA dataset

	Mask Ratio	**Vi-Trans without AFF**	**Vi-Trans with AFF**
PSNR ↑	0–20%	35.02	36.12
	20–40%	28.91	28.98
	40–60%	23.89	24.06
SSIM ↑	0–20%	0.958	0.963
	20–40%	0.821	0.893
	40–60%	0.887	0.888
LPIPS ↓	0–20%	0.021	0.015
	20–40%	0.070	0.051
	40–60%	0.132	0.127
FID ↓	0–20%	2.55	1.98
	20–40%	7.04	6.95
	40–60%	11.12	10.88

Table 6. Ablation study comparing Vi-Trans with and without AFF on Places2 dataset

	Mask Ratio	**Vi-Trans without AFF**	**Vi-Trans with AFF**
PSNR ↑	0–20%	31.98	32.12
	20–40%	28.12	28.88
	40–60%	22.98	22.99
SSIM ↑	0–20%	0.898	0.914
	20–40%	0.862	0.923
	40–60%	0.612	0.768
LPIPS ↓	0–20%	0.020	0.013
	20–40%	0.097	0.087
	40–60%	0.178	0.166
FID ↓	0–20%	10.23	9.99
	20–40%	12.89	11.88
	40–60%	13.91	13.55

5. Conclusion

In this work, we proposed Vi-Trans, a Vision Transformer-based autoencoder integrated with an Adaptive Feature Fusion (AFF) module for image inpainting. The

AFF module effectively combines the global contextual awareness of transformer layers with the local spatial precision of convolutional features, enabling accurate restoration of missing regions with enhanced texture realism, edge continuity, and structural coherence. Unlike conventional transformer-based inpainting methods that rely solely on self-attention, Vi-Trans dynamically balances global semantics and fine-grained details through adaptive fusion, producing visually coherent and perceptually rich reconstructions. The framework was further extended toward Metaverse applications such as avatar restoration, object removal, and visual continuity in VR/AR environments, demonstrating superior performance over state-of-the-art transformer-based models in both quantitative and perceptual metrics. Moreover, the proposed AFF mechanism offers a promising foundation for 3D and multi-view inpainting, where it can be adapted to fuse spatial and geometric information across multiple viewpoints. While current evaluations are limited to 2D datasets like CelebA, future work will focus on extending Vi-Trans to 3D and video-based scenarios by integrating depth cues and ensuring real-time performance, further advancing its applicability to immersive Metaverse environments.

Author contributions: Conceptualization, J.P.; methodology, J.P.; software, J.P.; formal analysis, J.P.; investigation, J.P.; data curation, J.P.; writing—original draft preparation, J.P.; visualization, J.P.; writing—review and editing, R.K.K.; supervision, R.K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Acknowledgment: The authors would like to thank their institution for providing the necessary facilities to carry out this research work.

Conflict of interest: The authors declare no conflict of interest.

References

1. Bertalmio M, Sapiro G, Caselles V, et al. Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques; July 2000; pp. 417–424. doi:10.1145/344779.344972.
2. Al-Jaberi AK, Hameed EM. A review of PDE-based local inpainting methods. Journal of Physics: Conference Series. 2021, 1818: 012149. doi:10.1088/1742-6596/1818/1/012149.
3. Jini P, Rajkumar KK. Image inpainting using image interpolation – an analysis. Revista Geintec-Gestão InovaçãoeTecnologias. 2021, 11(2): 1906–1920. doi:10.47059/revistageintec.v11i2.1807.
4. Jini P, Rajkumar KK. Application of Image Inpainting in Drug and Alcohol Addiction Research Using Repeated Convolution Method. Journal of Drug and Alcohol Research. 2024, 13: 236413. doi:10.4303/JDAR/236413.
5. Wang Y, Song B, Zhang Z. An image inpainting method based on generative adversarial networks inversion and autoencoder. IET Image Processing. 2024, 18(4): 1042–1052. doi:10.1049/iet-ipr.2023.0348.
6. Zhang, L., Yu, Y., Yao, J., & Fan, H. (2025). High-fidelity image inpainting with multimodal guided GAN inversion. International Journal of Computer Vision, 133(8), 5788-5805.
7. Deng Y, Hui S, Zhou S, et al. T-former: An efficient transformer for image inpainting. In: Proceedings of the 30th ACM International Conference on Multimedia (MM '22); October 2022; doi:10.1145/3503161.3548446.
8. Miao W, Wang L, Lu H, et al. ITrans: Generative image inpainting with transformers. Multimedia Systems. 2024, 30(1): 21. doi:10.1007/s00530-023-01211-w.
9. Xu Y, Lyu Y, Xiong G, et al. Adaptive feature fusion networks for origin-destination passenger flow prediction in metro systems. IEEE Transactions on Intelligent Transportation Systems. 2023, 24(5): 5296–5312. doi:10.1109/TITS.2023.3239101.
10. Buhalis D, Leung D, Lin M. Metaverse as a disruptive technology revolutionising tourism management and marketing. Tourism Management. 2023, 97: 104724. doi:10.1016/j.tourman.2023.104724.
11. Enamorado-Díaz E, García-García JA, Escalona-Cuaresma MJ, et al. Metaverse applications: Challenges, limitations and

opportunities — A systematic literature review. Information and Software Technology. 2025, 182: 107701. doi:10.1016/j.infsof.2025.107701.
12. Sun J, Gan W, Chao HC, et al. Metaverse: Survey, applications, security, and opportunities. arXiv preprint arXiv:2210.07990. 2022. doi:10.48550/arXiv.2210.07990.
13. Criminisi A, Pérez P, Toyama K. Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing. 2004, 13(9): 1200–1212. doi:10.1109/TIP.2004.833105.
14. Hays J, Efros AA. Scene completion using millions of photographs. ACM Transactions on Graphics. 2007, 26(3): 4. doi:10.1145/1276377.1276382.
15. Barnes C, Shechtman E, Finkelstein A, et al. PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics. 2009, 28(3): 24. doi:10.1145/1531326.1531330.
16. Pathak D, Krähenbühl P, Donahue J, et al. Context Encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; pp. 2536–2544. doi:10.1109/CVPR.2016.278.
17. Lizuka S, Simo-Serra E, Ishikawa H. Globally and locally consistent image completion. ACM Transactions on Graphics. 2017, 36(4): 107. doi:10.1145/3072959.3073659.
18. Liu, H., Jiang, B., Song, Y., Huang, W., & Yang, C. (2020).*Rethinking image inpainting via a mutual encoder–decoder with feature equalizations*. In Computer Vision – ECCV 2020 (European Conference on Computer Vision), Lecture Notes in Computer Science, vol. 12347, pp. 725–741. Springer.DOI: https://doi.org/10.1007/978-3-030-58536-5_43.
19. Shamsolmoali P, Zareapoor M, Granger E. TransInpaint: Transformer-based image inpainting with context adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2023; pp. 849–858. doi:10.1109/ICCV.2023.00080.
20. Liu G, Reda FA, Shih KJ, et al. Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018; pp. 85–100. doi:10.1007/978-3-030-01252-6_6.
21. Yu J, Lin Z, Yang J, et al. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019; pp. 4471–4480. doi:10.1109/ICCV.2019.00457.
22. Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122. 2015. doi:10.48550/arXiv.1511.07122.
23. Kumar V, Singh RS, Dua Y. Morphologically dilated convolutional neural network for hyperspectral image classification. Signal Processing: Image Communication. 2022, 101: 116549. doi:10.1016/j.image.2021.116549.
24. Braşoveanu AMP, Andonie R. Visualizing transformers for NLP: A brief survey. In: 2020 24th International Conference Information Visualisation (IV); September 2020; pp. 270–279. doi:10.1109/IV51561.2020.00051.
25. Chefer H, Gur S, Wolf L. Generic attention-model explainability for interpreting bi-modal and encoder–decoder transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021; pp. 397–406. doi:10.1109/ICCV48922.2021.00045.
26. Liu Z, Lin Y, Cao Y, et al. Swin Transformer: Hierarchical Vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021; pp. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
27. Chen Y, Xia R, Yang K, et al. DNNAM: Image inpainting algorithm via deep neural networks and attention mechanism. Applied Soft Computing. 2024, 154: 111392. doi:10.1016/j.asoc.2024.111392.
28. Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV); 2020; pp. 213–229. doi:10.1007/978-3-030-58452-8_13.
29. Csurka G, Volpi R, Chidlovskii B. Semantic image segmentation: Two decades of research. Foundations and Trends® in Computer Graphics and Vision. 2022, 14(1-2): 1–162. doi:10.1561/0600000095.
30. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems, NeurIPS 2017; 2017. doi:10.48550/arXiv.1706.03762.
31. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020. doi:10.48550/arXiv.2010.11929.
32. Touvron H, Cord M, El-Nouby A, et al. Three things everyone should know about vision transformers. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022; pp. 497–515. doi:10.1007/978-3-031-20053-3_29.
33. Liu R, Mi L, Chen Z. AFNet: Adaptive fusion network for remote sensing image semantic segmentation. IEEE Transactions

on Geoscience and Remote Sensing. 2021, 59(9): 7871–7886. doi:10.1109/TGRS.2020.3034123.

34. Li Y, Chen W, Huang X, et al. MFVNet: A deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation. Science China Information Sciences. 2023, 66(4): 140305. doi:10.1007/s11432-022-3599-y.
35. Li W, Lin Z, Zhou K, et al. Mat: Mask-aware transformer for large hole image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022; pp. 10758–10768. doi:10.1109/CVPR52688.2022.01049.
36. Huang W, Deng Y, Hui S, et al. Sparse self-attention transformer for image inpainting. Pattern Recognition. 2024, 145: 109897. doi:10.1016/j.patcog.2023.109897.
37. Wang Z, Bovik AC. Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Processing Magazine. 2009, 26(1): 98–117. doi:10.1109/MSP.2008.930649.
38. Lucas A, Lopez-Tapia S, Molina R, et al. Generative adversarial networks and perceptual losses for video super-resolution. IEEE Transactions on Image Processing. 2019, 28(7): 3312–3327. doi:10.1109/TIP.2019.2895768.
39. Rojas DJB, Fernandes BJT, Fernandes SMM. A review on image inpainting techniques and datasets. In: Proceedings of the 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI); 2020; pp. 240–247. doi:10.1109/SIBGRAPI51738.2020.00040.
40. Setiadi DRIM. PSNR vs SSIM: Imperceptibility quality assessment for image steganography. Multimedia Tools and Applications. 2021, 80(6): 8423–8444. doi:10.1007/s11042-020-10035-z.
41. Jayasumana S, Ramalingam S, Veit A, et al. Rethinking FID: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024; pp. 9307–9315. doi:10.1109/CVPR52733.2024.00889.
42. Chen Z, Wang X, Xie L, et al. LPIPS-AttnWav2Lip: Generic audio-driven lip synchronization for talking head generation in the wild. Speech Communication. 2024, 157: 103028. doi:10.1016/j.specom.2024.103028.
43. Wan, Z., Zhang, J., Chen, D., & Liao, J. (2021). High-fidelity pluralistic image completion with transformers. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4692-4701).