HINT: High-quality INpainting Transformer with Mask-Aware Encoding and Enhanced Attention

Shuang Chen, Amir Atapour-Abarghouei, Hubert P. H. Shum†, *Senior Member, IEEE*

***Abstract*—Existing image inpainting methods leverage convolution-based downsampling approaches to reduce spatial dimensions. This may result in information loss from corrupted images where the available information is inherently sparse, especially for the scenario of large missing regions. Recent advances in self-attention mechanisms within transformers have led to significant improvements in many computer vision tasks including inpainting. However, limited by the computational costs, existing methods cannot fully exploit the efficacy of long-range modelling capabilities of such models. In this paper, we propose an end-to-end High-quality INpainting Transformer, abbreviated as HINT, which consists of a novel mask-aware pixel-shuffle downsampling module (MPD) to preserve the visible information extracted from the corrupted image while maintaining the integrity of the information available for high-level inferences made within the model. Moreover, we propose a Spatially-activated Channel Attention Layer (SCAL), an efficient self-attention mechanism interpreting spatial awareness to model the corrupted image at multiple scales. To further enhance the effectiveness of SCAL, motivated by recent advanced in speech recognition, we introduce a sandwich structure that places feed-forward networks before and after the SCAL module. We demonstrate the superior performance of HINT compared to contemporary state-of-the-art models on four datasets, CelebA, CelebA-HQ, Places2, and Dunhuang.**

***Index Terms*—Image Inpainting, Transformer, Representation Learning**

I. Introduction

IMAGE inpainting is a computer vision task that aims to reconstruct an image based on the visible pixels of a damaged or corrupted image with missing regions. Its applications span across image processing and computer vision tasks such as photo editing [1], objective removal [2] and depth completion [3].

Image inpainting has been greatly benefited by modern learning-based techniques [4]–[12], even though it has existed long before the widespread use of deep learning [13]–[15]. Many existing image inpainting methods [16], [17] with Convolutional Neural Networks (CNN) commonly use an encoder-decoder architecture, which down-sample the corrupted image to a hidden latent space, and then up-sample it to produce a restored image with comparable semantics and structure to the original [4], [5]. These CNN-based methods typically use local convolutional filters, which possess a limited receptive field and solely capture information within a restricted local region.

Manuscript created august, 2023.

S. Chen, A. Atapour-Abarghouei and H. P. H. Shum are with Durham University, UK. (e-mail: {shuang.chen, amir.atapour-abarghouei, hubert.shum}@durham.ac.uk).

†Corresponding author: H. P. H. Shum

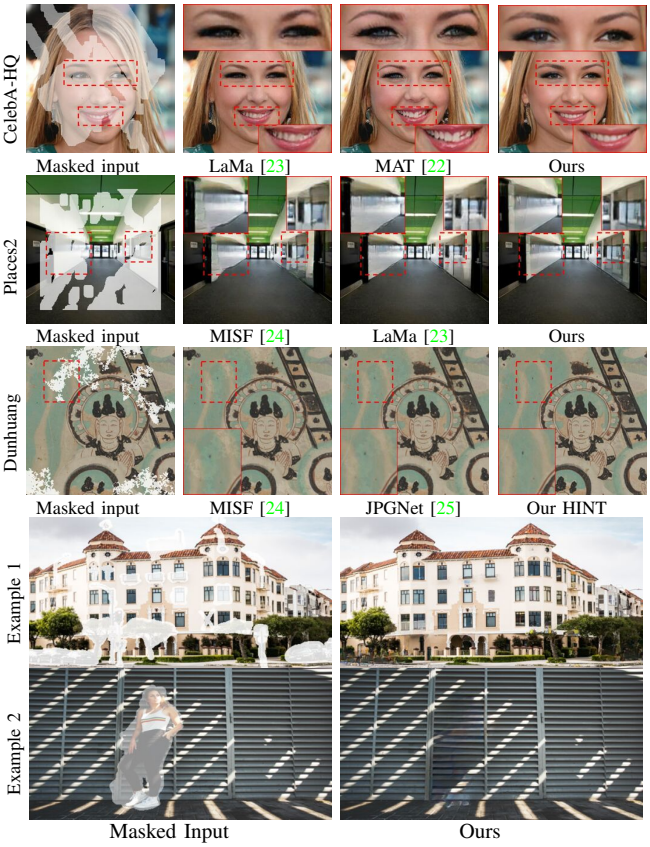

Fig. 1. Comparisons with the state of the art [22]–[24] on different datasets [26]–[28] with large masks (shown in white areas). Red boxes highlight major differences. The bottom two examples are from unseen real-world high-resolution images.

These methods suffer from the limitation to capture long-range spatial relations between distant image regions, thus compromising their effectiveness in image inpainting [4], [6], [18]. To address this challenge, [19], [20] incorporate spatial self-attention into the network, which involves calculating a self-attention map in the deep latent space to capture long-range dependencies between feature elements. Recent work attempted to introduce transformer architectures to image inpainting [20]–[22], which enable global long-range modeling of images that have been down-sampled or partitioned into patches, resulting in improved performance.

A significant challenge hindering image inpainting is effectively modeling the valid information within visible regions, which is crucial for reconstructing semantically coherent and texture-consistent details in the missing regions. This is particularly noticeable in large masked regions, where the valid information is limited. Existing methods that utilise

convolutional layers for downsampling come with the inherent drawback of information loss [29], attributed to the reduction of feature size from filters and downsampling. Given its capability to preserve input information, pixel-shuffle down-sample is widely used in image denoising [30], image deraining [31] and image super-resolution [32]. It periodically rearranges the elements of the input into an output scaled by the sample stride. However, its effectiveness depends on the assumption that the sample stride is small enough to avoid disrupting the noise distribution [33]. This holds only for a relatively independent distribution of raindrops and noise, and is not suitable for image inpainting with irregular and variable-size masks. Simply using conventional Pixel-shuffle Down-sampling (PD) [30]–[32] for corrupted image would lead to the problem of pixel drifting, which is shown in Fig. 3 (upper branch). The pixel drifting happens in $\hat{X}$. After the feature X' is downsampled, the position of the masked regions (white elements) becomes inconsistent across channels, causing the visible area to be misaligned in the channel, disrupting subsequent feature extraction processes within the model, thus affecting the accurate modelling of the valid information from the visible regions of the input image.

Another challenge in applying spatial self-attention in CNN-based models is its significant computational expense. Considering this, spatial self-attention is typically only employed on low-resolution representations [19], [20]. While transformer-based methods [21], [22] employ multiple spatial self-attention blocks to model long-range dependencies. However, the quadratic computational complexity limits their wider applicability. To address this, the prevalent compromise involves down-sampling [22] or reducing the resolution [21] of the input image prior to being passed through the transformer. However, this strategy leads to information loss from the input images through the model, which is detrimental to image inpainting where visible information is already limited. This loss subsequently results in the degradation of fine-grained features. As long-range dependencies are modelled over these degraded features, the reconstructed output suffer from blurring artefacts and vague structures. [20], [21] introduce extra refinement networks to improve image quality after getting coarse completed images, rather than recovering high-quality results directly. The method in [34] replaces spatial self-attention with channel self-attention to reduce computational complexity. Although channel self-attention gains linear computational complexity, it completely loses spatial awareness. This makes it possible to highlight "what" the salient features are but cannot discern "where" the spatially important regions are, which is essential as visible regions often exhibit complex and irregular shapes, especially with large irregular masks. Some existing works [35]–[37] attempt to address the spatial awareness loss by incorporating spatial self-attention back to the channel self-attention, but at a cost of significant increases in computation.

To address these common challenges currently restricting progress in the existing literature, we present a novel High-quality INpainting Transformer (HINT) for image inpainting, which enables efficient multiscale modeling of the global context while minimising the loss of valid information. Specifically, we propose a tailor-made pixel-shuffle down-sampling (MPD) module for image inpainting to reduce information loss and maintain the consistency of data. To enhance the representation learning capabilities of our model, we develop a Spatially-activated Channel Attention Layer (SCAL) to blend information in both the channel and spatial dimensions. Unlike these existing methods [35]–[37], the innovation of SCAL lies in its minimalistic and efficient design, only utilising convolutional layers to retrain spatial awareness, thereby mitigating the significant computational cost, which is a major issue in the field. This enhanced self-attention module plays the predominant role in HINT and build HINT as a transformer-based model. To further improve the effectiveness of SCAL with limited parameters, we propose a module known as the "Sandwich", sandwiching the proposed SCAL between two feed-forward networks (FFNs) for each transformer block. This structure results in better performance compared to alternative designs with the same number of network parameters.

Comparative experiments show that HINT outperforms state-of-the-art image inpainting approaches (Fig. 5) across four datasets, i.e., CelebA [38], CelebA-HQ [26], Places2 [27] and Dunhuang challenge [28]. We also perform ablation experiments to demonstrate the contribution of proposed components in HINT.

Our source code is openly released at https://github.com/ChrisChen1023/HINT.

Our major contributions are as follows:

- We propose HINT, an end-to-end transformer-based architecture for image inpainting that takes advantage of multi-scale feature- and spatial-level representations as well as pixel-level visual information.
- We propose a plug-and-play mask-aware pixel-shuffle down-sampling (MPD) module to preserve useful information while keeping irregular masks consistent during downsampling (Section III-B).
- We propose a Spatially-activated Channel Attention Layer (SCAL) using self-attention and convolutional attention to sequentially refine features at the channel and spatial dimensions. We further design an improved sandwich-shaped transformer block to boost the efficacy of the proposed SCAL (Section III-C).

II. Related Work

We consider prior work within two distinct areas: image inpainting (Section II-A), and visual transformers (Section II-B), which have gained prominence as effective techniques for addressing the image inpainting task.

A. Image Inpainting

Image inpainting predates learning-based techniques and the literature on image completion based on conventional strategies is extensive. Diffusion-based approaches complete minor and narrow stretches using neighbouring visible pixels [13]. Exemplar-based methods infer missing regions with plausible edge information based on other patches from background or external data [14], [15]. However, despite their ability to

plausibly reconstruct images with small and constrained missing regions, these methods are not fully capable of generating innovative features if they do not already exist in the known regions of images.

Compared with traditional methods, learning-based methods have achieved great success in inpainting, especially when it comes to generating new contextually sound content for large missing regions. [4] proposed a parametric framework for image inpainting based on an encoder-decoder architecture taking advantage of a Generative Adversarial Network (GAN) [39]. Subsequently, numerous GAN-based methods emerged to offer improved inpainting quality [6]–[8], [10], [18], [40]–[42] using better training strategies.

[5] use two discriminators to calculate both global and local adversarial losses. [43] propose region-wise normalisation for missing and visible areas. Partial [6] and gated convolutions [7], [8] are introduced to handle the irregular masks by improving the convolution operation [40], [41] to efficiently extract valid information for inpainting. [9] propose contextual attention to facilitate the matching of feature patches across distant spatial locations. Building on this, [10], [11] extended [9] by incorporating a multi-scale patch size to further improve its efficiency. [44] introduces fourier convolution-based encoder for image inpainting to avoid generating invalid feature inside the missing regions. These strategies all aim to explore how to effectively extract valid information from known regions for hole-filling, but they still suffer from the information loss caused by convolutional downsampling.

B. Visual Transformers

The notable success of Transformers [45] in natural language processing has recently prompted research into their applicability in computer vision [46], [47]. Driven by this, efforts were focused towards applying transformers to image inpainting [20]–[22], [48]–[50]. However, spatial-based self-attention incurs an expensive computational cost. To reduce computation, [21], [48] down-sample the input image into a lower resolution. [20], [22], [50] calculate the spatial self-attention after encoding the input image into low-resolution features. Nonetheless, these approaches fail to change the quadratic complexity of spatial self-attention, which restricts its applicability to high-frequency features.

Swin Transformer [46] reduces the computational complexity to linearity. However, the shifted-window design splits the local neighbourhood context of the visible and missing area, and thus is not ideally suited for inpainting. [34] propose utilising channel-wise self-attention in multi-scale representation with linear complexity for image reconstruction. Its variant [51] demonstrates the applicability in image inpainting. Nevertheless, both of these models omit spatial attention that is vital in delivering high-quality and contextually sound results. In contrast, our model integrates multiscale channel and spatial attention in an efficient manner, thus resolving the issue that prior work has struggled with [20]–[22].

III. HINT: High-quality INpainting Transformer

Formally, the problem is formulated as follows: the input image, I_{input}, is obtained by concatenating masked image, $I_M = I \odot M$, and the mask, M. The input image, I_{input}, is then processed by our proposed HINT model and a semantically accurate output image, I_C, will be generated. The whole formulation is denoted as: $I_C = HINT(I_{input})$.

We present our transformer-based HINT approach to image inpainting, which takes advantage of our novel Mask-aware Pixel-shuffle Down-sampling (MPD) to solve the information loss issue during downsampling and further enhance the use of valid information from known areas. Within the architecture, we propose a Spatially-activated Channel Attention Layer (SCAL), which aims to handle spatial awareness while maintaining efficiency within the transformer block. The SCAL is encapsulated between two feed-forward networks, forming a sandwich-shaped transformer block, henceforth referred to as "*Sandwich*". This design enables the effective extraction of long-range dependencies while preserving the smooth and coherent flow of valid information through the model.

A. The Overall Pipeline

Overall, as seen in Fig. 2, HINT consists of an end-to-end network with a gated embedding layer to selectively extract features, followed by a transformer body for modelling long-range correlations, and a projection layer to generate the output. Specifically, we insert a gating mechanism [7] into the embedding layer serving as a feature extractor, achieved by using two parallel paths of vanilla convolutions with one path activated by a GELU non-linearity [52] to dynamically embed the finer-grained features, leading to stronger representation learning and better optimisation [53]. The transformer body is an encoder-decoder architecture comprising multiple transformer blocks. The encoder consists of the first three blocks, each followed by an MPD layer to mitigate incoherence in invalid locations, while the final three blocks with conventional pixel shuffle upsampling form the decoder. Mirrored blocks are connected via skip connections to preserve shared features learned within the encoder. At the end, a convolutional layer is used to project the decoded features to the final output.

B. Mask-aware Pixel-shuffle Down-sampling

Conventional Pixel-shuffling Down-sampling (PD) is the inverse operation of Pixel-shuffle [54]. It periodically rearranges the input $T_{in} \in \mathbb{R}^{H \times W \times C}$ into $T_{out} \in \mathbb{R}^{\frac{H}{s} \times \frac{W}{s} \times s^2 C}$ for downsampling with s being the scale factor to denote the sample stride. PD can effectively preserve the input information, which is desirable for inpainting, particularly for reconstructing high-quality images. However, as PD uses non-overlapping sampling with stride s to generate mosaics from the image [54], the consistency of missing pixel locations can be disrupted during the down-sampling, as shown in Fig. 3, making it unsuitable for image inpainting.

We propose a Mask-aware Pixel-shuffle Down-sampling (MPD) module, which is a novel down-sampling approach specifically tailored for image inpainting. It resolves the issue of positional drift of masked pixels that occurs during the process of conventional PD. Furthermore, in contrast to convolution-downsampling, MPD preserves all valid information, thereby minimising information loss. Apart from

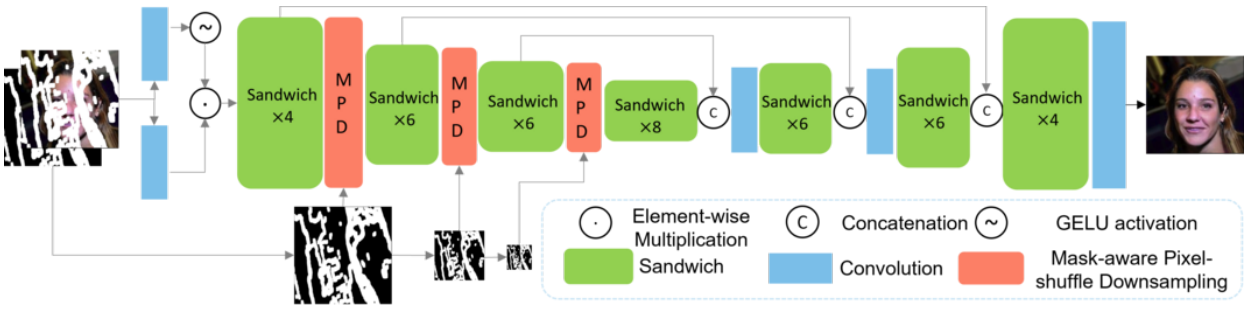

Fig. 2. The overview of the proposed framework, which is built with a gated embedding block, with multiple stacked "sandwiches" in different levels. The "sandwich" is described in Sec. III-C2, the MPD is described in Sec. III-B

inpainting, this module can be plugged into any other problem that involves masking, such as any that might use image segmentation labels masks as their input.

Given the features $X \in \mathbb{R}^{H\times W\times C}$ and mask $M \in \mathbb{R}^{H\times W\times 1}$, we first project X into X' with half the channels but the same size [54], utilising a 3×3 convolution operator $h(\cdot)$, and perform PD on both X' and M:

$$\hat{M} = PD(M), \hat{X} = PD(h(X)). \tag{1}$$

As shown in Fig. 3, the positions of the missing pixels in $\hat{X}$ drift and are discontinuous across channels while each channel of $\hat{M}$ sequentially indicates the positions of valid and invalid pixels in $\hat{X}$. To enforce $\hat{M}$ to act on the corresponding channel accurately, we intersperse and concatenate the sliced $\hat{X}$ and $\hat{M}$ across the channel, obtaining $\hat{X}_c \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 2C}$.

$$\hat{M}^0, \hat{M}^1, \hat{M}^2, \hat{M}^3 = Slice(\hat{M}), \tag{2}$$

$$\begin{aligned}\hat{X}^0, \hat{X}^1, \hat{X}^2, \ldots, \hat{X}^{2C-1} &= Slice(\hat{X}),\\ \hat{X}_c = &(\hat{X}^0||\hat{M}^0)||\ldots||(\hat{X}^i||\hat{M}^{(i+4)\%4})||\ldots\\ &||(\hat{X}^{2C-1}||\hat{M}^3),\end{aligned} \tag{3}$$

where $Slice(\cdot)$ is a channel-wise slice, $||$ is channel-wise concatenation, and $\%$ denotes the modulo operator. Thus, each feature has a paired mask as an indicator. In the end, we exploit a separable convolutional layer [55], denoted as $\phi(\cdot)$, to encode pairs of features and masks, aiming to learn the correct local priors from the features indicated by the shuffled mask, and forcing the encoder to accurately model the valid information within the visible regions. The output is formulated as:

$$X_{out} = \phi(\hat{X}_c). \tag{4}$$

C. The Transformer Body

Each of the seven transformer blocks stacks multiple *sandwiches* encapsulating the proposed SCAL for local-global representation learning, working with MPD to down-sample the features and control data flow consistency (Fig. 2).

1) Spatially-activated Channel Attention Layer: We propose a Spatially-activated Channel Attention Layer (SCAL) to strengthen the model to capture inter-channel dependencies while preserving spatial awareness. Channel self-attention [56] is computationally viable for high-resolution features due to its linear time and memory complexity growth with channel depth. However, it fails to account for "where" the important information is across the entire spatial position, thus ignoring the relationship between feature patches. This is very important for image inpainting as the global context in the valid regions within each image can be distinct and irregularly shaped, as defined by the irregular mask M.

To alleviate this issue, we improve the concept of transposed attention [34] by introducing a convolution-attention branch to capture the attention matrix of spatial locations. This enables HINT to effectively model long-range dependencies in the channel dimension, while attending to spatial locations where features should be emphasised. Unlike alternative approaches [20]–[22], [46], [47], SCAL does not increase the computational cost quadratically with input resolution, making it feasible for multi-scale context modelling.

As shown in Fig. 4, SCAL contains two branches. Given input feature X, the channel self-attention branch is:

$$\begin{aligned}X_c &= LN(X),\\ \hat{X}_c &= \left(W^V_{d3}W^V_1 X_c\right)\cdot \mathrm{Attc}(X_c),\\ \mathrm{Attc}(X_c) &= \varphi\left(\frac{W^Q_{d3}W^Q_1 X_c \cdot \left(W^K_{d3}W^K_1 X_c\right)^T}{\gamma}\right),\end{aligned} \tag{5}$$

where *LN* denotes layer normalisation, γ is a learnable parameter to scale the dot product of key and query, W_1 is the linear projection and W_{d3} is the 3×3 depth-wise convolution, $\mathrm{Attc}(\cdot)$ represents the function to calculate the channel attention map, and φ is a softmax layer. In the spatial branch, we first down-sample the input features X but not fully squeeze, via average pooling to preserve global spatial information. Subsequently, two 3×3 convolutions serve as attention descriptors followed by an upsampling process, generating a soft global attention matrix, $\alpha = \mathrm{Atts}(X)$, which is used to reweight the output obtained through channel attention:

$$\mathrm{Atts}(X) = \mathrm{Up}(f(g(\mathrm{AP}(X)))), \tag{6}$$

where AP is an average pooling layer, Up is upsampling. $f(\cdot)$ and $g(\cdot)$ are two similar convolution blocks, one of which contains a 3×3 convolutional layer, a normalisation layer, and a ReLU layer [57]. $\mathrm{Atts}(\cdot)$ represents the function to calculate the spatial attention map. As depicted in Fig. 4, the attention matrix α modulates the output of the channel branch $\hat{X}_c$ through point-wise multiplication. Subsequently, the mapping function $\theta(\cdot)$ is a projection layer performed via

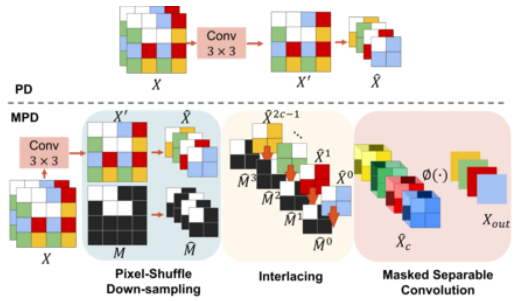

Fig. 3. The comparison of Pixel-shuffle Down-sampling (PD, upper) and the proposed Mask-aware Pixel-shuffle Down-sampling (MPD, lower). Ours proposed MPD, with one 3×3 convolution, a conventional PD, interlacing (concatenation of feature and mask slices), and a masked-separable convolution. Invalid pixel drifting happens in $\hat{X}$. After the feature X' is downsampled, the masked position becomes inconsistent across channels.

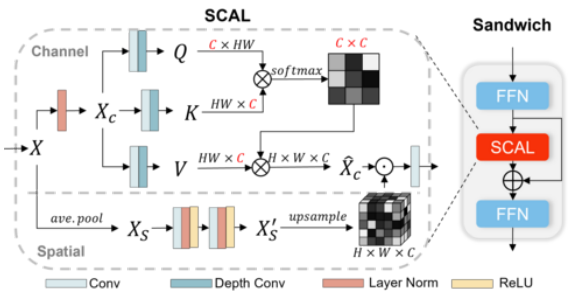

Fig. 4. "Sandwich" (right) and "Spatially-activated Channel Attention Layer" (left). "$\oplus$","$\otimes$", and "$\odot$" denote the element-wise sum, matrix multiplication, and element-wise multiplication, respectively.

of a 1×1 convolution. The complete representation of the SCAL is:

$$SCAL(X) = \theta(\hat{X}_c \odot \text{Atts}(X)). \tag{7}$$

2) Sandwich-shaped Transformer Block: Image inpainting presents a significant challenge: the network must effectively learn from limited context to reconstruct complete images. This task is particularly daunting when faced with irregularly shaped masks, which complicate feature extraction, especially in areas with extensive missing information. This process of masking in image inpainting bears a notable resemblance to the masking of audio spectrograms in speech recognition for data augmentation purposes, as seen in techniques like SpecAugment [58], [59]. The Conformer [60], with its innovative "FFN-Attention-Conv-FFN" architecture, demonstrates remarkable efficiency in speech recognition by using augmented, masked spectrograms as inputs. We hypothesise that such structures are equally effective for image inpainting, since their inputs are also incomplete and insufficient, highlighting a common challenge in both fields that may benefit from similar architectural solutions.

Therefore, to boost the effectiveness of our attention layer, we propose a sandwich-shaped transformer block with an FFN-Attention-FFN structure. This first FFN serves as a filter, extracting more essential features for the following attention layer to capture long-distance dependencies (see Section IV-D for validations). Unlike [60], we remove the convolutional layer in the middle, and enhance the two FFNs with depth-wise convolutions with a gate mechanism [34]. This is because FFN integrating depth-wise convolution captures local information from every channel, which helps the model learn a more comprehensive and informative feature representation with fewer parameters [55]. Also, the gating strategy selectively filters and modulates the information flow according to the importance of each feature to the final high-quality output, thereby reducing irrelevant information and highlighting the most salient input features for representation learning. Given an input $X \in \mathbb{R}^{H \times W \times C}$, our sandwich is formulated as:

$$X_{\text{out}} = FFN(SCAL(FFN(X))). \tag{8}$$

D. Loss Functions

To obtain high-quality inpainting results, we follow the established literature [18], [24] to develop multiple loss components, including an $\mathcal{L}_1$ loss to enforce a contextually sound reconstruction, style loss $\mathcal{L}_{\text{style}}$ to measure the difference in style, perceptual loss $\mathcal{L}_{\text{perc}}$ to compare the high-level perceptual features extracted from a pre-trained network, and an adversarial loss $\mathcal{L}_{\text{adv}}$ to improve overall output quality. The final loss function is thus denoted as:

$$\begin{aligned}\mathcal{L}_{total}(\hat{\mathbf{I}}, \mathbf{I_{gt}}) =& \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{\text{style}} + \lambda_3 \mathcal{L}_{\text{perc}} \\ &+ \lambda_4 \mathcal{L}_{\text{adv}},\end{aligned} \tag{9}$$

where the weighting coefficients $\lambda_1 = 1$, $\lambda_2 = 250$, $\lambda_3 = 0.1$, $\lambda_4 = 0.001$ were chosen based on the parameter analysis (see Section IV-D).

IV. Experiments

In this section, we present a comprehensive evaluation of the proposed HINT. First, we describe the datasets employed and delve into the specifics of the implementation. Then, we compare HINT with state-of-the-art methods to showcase its superior performance, with both quantitative and qualitative results. Finally, we conduct thorough ablation studies to evaluate the significance of each proposed component.

A. Datasets

To assess the efficacy of our proposed method, we employ CelebA [38], CelebA-HQ [26], Places2-Standard [27] and Dunhuang Challenge [28] datasets. All experiments are conducted with 256×256 images, providing a comprehensive evaluation of our approach in a consistent and well-defined setting. The CelebA [38] and CelebA-HQ [26] are two human face datasets with different qualities, while the Places2-Standard dataset is a subset of the Places2 [27] dataset offering a diverse collection of scenes, such as indoor and outdoor environments, natural landscapes, and man-made structures and constructions. These three datasets are commonly used within the existing literature on inpainting [20]–[22], making them ideal for evaluating our approach. The Dunhuang Challenge [28] dataset represents a practical application of image inpainting in real-world scenarios.

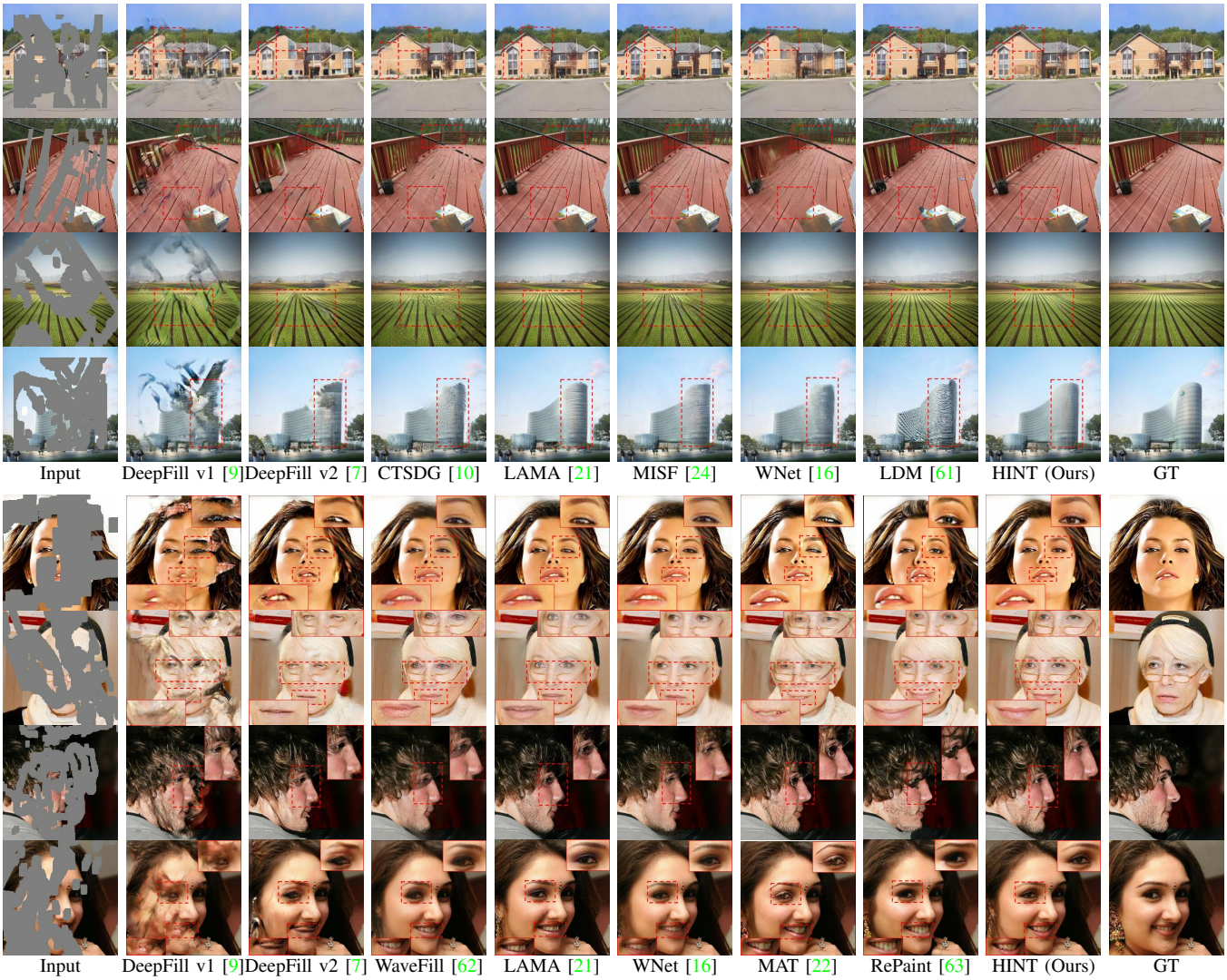

Fig. 5. Comparisons with visualisations (256×256) showing that our results are more coherent in structure and sharper in texture and semantic details. The top two rows are from CelebA-HQ [26] and the bottom two rows are from Places2 [27].

For CelebA and Dunhuang, we follow the standard configuration to split the data for training and testing. In the case of the CelebA-HQ dataset, to ensure reproducibility, we use the first 28,000 images for training and the remaining 2,000 images for testing. For the Places2-Standard dataset, we use the standard training set and validation set for training and testing, respectively. For mask settings, we follow prior work [10], [24] and use irregular masks [6] for CelebA, CelebA-HQ, and Places2. As for Dunhuang Challenge, we use the officially released masks for testing.

B. Implementation Details

In the 7-level transformer blocks, the number of Sandwich blocks is sequentially set to [4,6,6,8,6,6,4] and the attention head in SCAL are [1,2,4,8,4,2,1]. All experiments are carried out on a single NVidia A100 GPU with a batch size of 4. We adopt the Adam optimiser [65] with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is initially set to $1e^{-4}$ and is halved at the 75% milestone of the training progress. Compared to the state of the art in the existing literature [10], [21], [66], our approach is more robust against small changes in the training procedure, making it more generalisable and easier to deploy. Our training pipeline does not rely on warm-up step [21], pre-training requirements [66] or fine-tuning [10].

C. Comparison with the State of the Art

In assessing our HINT, designed to generate high-quality, fine-grained images, we follow [24] to employ a suite of evaluation metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), L1 and Perceptual Similarity (LPIPS). These chosen metrics align with our intent to create a nuanced and comprehensive understanding of the performance of models. PSNR and L1 are used to measure pixel-wise reconstruction accuracy, which reflects the fidelity of the inpainted output. SSIM [67] evaluates structural similarity, ensuring the inpainted segments remain coherent within the image contextually. We also include LPIPS [68], a learned perceptual metric, capable of detecting complex distortions that mirror human perceptual differences, a crucial attribute when the aim is to produce high-quality imagery.

TABLE I
COMPARISON RESULTS ON (A, TOP) CELEBA-HQ, (B, MIDDLE) CELEBA AND (C, BOTTOM) PLACES2. THE **BOLD** AND UNDERLINE INDICATE THE BEST AND THE SECOND BEST RESPECTIVELY.

CelebA-HQ	0.01%-20%					20%-40%					40%-60%				
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [9]	34.2507	0.9047	1.7433	2.2141	0.1184	26.8796	0.8271	2.3117	9.4047	0.1329	21.4721	0.7492	4.6285	15.4731	0.2521
DeepFill v2 [7]	34.4735	0.9533	0.5211	1.4374	0.0429	27.3298	0.8657	1.7687	5.5498	0.1064	22.6937	0.7962	3.2721	8.8673	0.1739
LaMa [23]	35.5656	0.9685	0.4029	1.4309	0.0319	28.0348	0.8983	1.3722	4.4295	0.0903	23.9419	0.8003	2.8646	8.4538	0.1620
WNet [16]	35.3591	0.9647	0.4957	1.2759	0.0287	28.1736	0.8872	1.4495	4.7299	0.0833	23.8357	0.7872	2.9316	9.4926	0.1649
MAT [22]	35.5466	0.9689	0.3961	1.2428	0.0268	27.6684	0.8957	1.3852	3.4677	0.0832	23.3371	0.7964	2.9816	5.7284	0.1575
WaveFill [62]	31.4695	0.9290	1.3228	6.0638	0.0802	27.1073	0.8668	2.1159	8.3804	0.1231	23.3569	0.7817	3.5617	13.0849	0.1917
Ours	**36.5725**	**0.9777**	**0.3942**	**1.1128**	**0.0228**	**28.6247**	**0.9195**	**1.2885**	**3.3915**	**0.0745**	**24.1287**	**0.8241**	**2.7778**	**5.6179**	**0.1449**

Places2	0.01%-20%					20%-40%					40%-60%				
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
DeepFill v1 [9]	30.2958	0.9532	0.6953	26.3275	0.0497	24.2983	0.8426	2.4927	31.4296	0.1472	19.3751	0.6473	5.2092	46.4936	0.3145
DeepFill v2 [7]	31.4725	0.9558	0.6632	23.6854	0.0446	24.7247	0.8572	2.2453	27.3259	0.1362	19.7563	0.6742	4.9284	36.5458	0.2891
CTSDG [10]	32.111	0.9565	0.6216	24.9852	0.0458	24.6502	0.8536	2.1210	29.2158	0.1429	20.2962	0.7012	4.6870	37.4251	0.2712
WNet [16]	32.3276	0.9372	0.5913	20.4925	0.0387	25.2198	0.8617	2.0765	24.7436	0.1136	20.4375	0.6727	4.6371	32.6729	0.2416
MISF [24]	32.9873	0.9615	0.5931	21.7526	0.0357	25.3843	0.8681	1.9460	30.5499	0.1183	20.7260	0.7187	4.4383	44.4778	0.2278
LaMa [23]	32.4660	0.9584	0.5969	14.7288	0.0354	25.0921	0.8635	2.0048	22.9381	0.1079	20.6796	0.7245	4.4060	25.9436	0.2124
WaveFill [62]	29.8598	0.9468	0.9008	30.4259	0.0519	23.9875	0.8395	2.5329	39.8519	0.1365	18.4017	0.6130	7.1015	56.7527	0.3395
Ours	**33.0276**	**0.9689**	**0.5612**	**13.9128**	**0.0307**	**25.4216**	**0.8807**	**1.9270**	**20.0241**	**0.1003**	**20.9243**	**0.7470**	**4.3296**	**25.7150**	**0.2041**

CelebA	0.01%-20%					20%-40%					40%-60%				
Method	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
CTSDG [10]	36.465	0.9732	0.5871	2.5876	0.0334	29.1393	0.9159	1.38	7.4925	0.0935	23.8371	0.8157	3.04	9.8473	0.1815
MISF [24]	36.8981	0.9747	0.3441	3.3598	0.0333	28.9270	0.9103	1.227	8.0249	0.1031	23.5355	0.8033	3.182	13.2475	0.2012
Ours	**37.5696**	**0.9754**	**0.3402**	**1.0270**	**0.0232**	**29.8525**	**0.9208**	**1.220**	**4.1359**	**0.0689**	**24.4538**	**0.8270**	**2.7802**	**5.3612**	**0.1408**

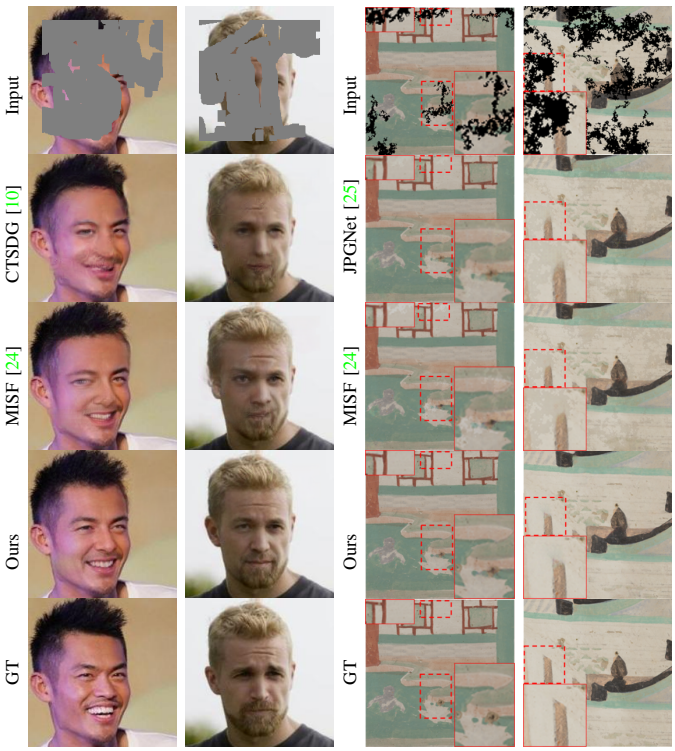

Fig. 6. Comparisons with SOTAs [10], [24], [25] on CelebA [38] (masks shown in grey) and Dunhuang [28] (masks shown in black). Red boxes highlight differences.

TABLE II
COMPARISONS ON THE DUNHUANG CHALLENGE DATASET.

Model	PSNR↑	SSIM↑	L1↓	LPIPS↓
StructFlow [64]	35.199	0.9559	0.475	0.0589
EdgeConnect [18]	36.419	0.9635	0.441	0.0480
RFRNet [40]	36.485	0.9648	0.401	0.0463
JPGNet [25]	37.646	0.9724	0.353	0.0469
MISF [24]	38.383	0.9735	0.341	0.0330
Ours	**38.6705**	**0.9743**	**0.3161**	**0.0286**

TABLE III
NUMBER OF PARAMETER AND INFERENCE TIME

Model	Param $\times 10^6$	Infer. Time/per img
DeepFill v1 [9]	3	7 ms
DeepFill v2 [7]	4	10 ms
Wavefill [62]	49	70 ms
CTSDG [10]	52	20 ms
WNet [16]	46	35 ms
MISF [24]	26	10 ms
MAT [22]	62	70 ms
LAMA [23]	51	25 ms
Stable Diffusion	860	880 ms
LDM [61]	387	6000 ms
Repaint [63]	552	250000 ms
Ours	139	125 ms

We categorise the masks into three groups based on the mask ratio, i.e., small (0.01%-20%), medium (20%-40%) and large (40%-60%), referring to the extent of missing regions.

Quantitative Results As shown in Tab. I and Tab. II, HINT achieves a better overall performance across all datasets and mask ratios than the state of the arts [10], [21], [22], [24], [25], [62]. Compare to the latest transformer-based MAT [22] on CelebA-HQ, HINT improves PSNR by 5.7%, 3.3% and 3.4% at the increasing mask ratios respectively, demonstrating that it preserves more high-fidelity details in reconstructed images. In Places2, compared with the latest high-quality inpainting method MISF [24], HINT achieves a 12.6%, 13.8% and 7.2% decrease for LPIPS, showcasing its effectiveness in perceptual recovery. Since the Dunhuang Challenge provides standard masks, we crawled the benchmark from [24] for comparison. HINT outperforms existing models across all metrics.

For a comprehensive and robust evaluation, we also compare our model with the state-of-the-art diffusion model-

TABLE IV
COMPARISON WITH DIFFUSION MODELS.

Plces2	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
LDM [61]	19.6476	0.7052	4.6895	27.3619	0.2675
Stable Diffusion*	19.4812	0.7185	4.5729	27.8830	0.2416
Ours	**20.8579**	**0.7227**	**4.3814**	**26.7895**	**0.2102**
CelebA-HQ	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
RePaint [63]	21.8321	0.7791	3.9427	8.9637	0.1943
Ours	**24.1287**	**0.8241**	**2.778**	**7.5793**	**0.1449**

*: The officially released Stable Diffusion inpainting model pretrained on high-quality LAION-Aesthetics V2 5+ dataset.

based methods with large masks, which are well-known for their prowess in generating high-quality images [69]. Three prominent diffusion models, LDM [61], Stable Diffusion (SD) and RePaint [63], are chosen for comparison. To allow for a fair comparison, all experiments are conducted on officially released pretrained models on the corresponding datasets. It is important to note that SD does not provide models pretrained on either CelebA-HQ or Places2, so, we chose the LAION v2 5+ pre-trained model, as its data distribution is similar to that of the Places2 dataset, but it is much larger and of higher quality. Tab. IV and Tab. III underscore the superior performance of our model across all metrics and signify the efficiency in image inpainting tasks. Ideally, we wish to assess all diffusion model on Places2. However, due to the significant inference time required by RePaint (Tab. III), a single evaluation on the Places2 dataset for three mask ratios demands around one GPU-year, making it computationally intractable. As a result, we chose to evaluate LDM on Places2, given its relatively more manageable inference time, and focused our analysis of RePaint on the CelebA-HQ dataset.

Qualitative Results We provide the exemplar visual results to further demonstrate the advantages of HINT over comparators. As shown in Fig. 5, our model generates high-quality images with more coherent structures and fewer artifacts, such as roofs and planks. For face restoration, our model better recovers finer-grained details, such as eye features, compared to the current state of the art [21], [22], [62]. We also provide qualitative results for CelebA [38] and Dunhuang datasets [28] in Fig. 6 to indicate our superior performance in global context modeling. The proposed HINT recovers high-quality faces with clear textures and plausible semantics, even with a large mask covering almost all facial attributes. The results on Dunhuang show that our model suppresses the generation of light mottle, and demonstrates the effectiveness of our model in handling small scratch masks.

Efficiency Comparison Our model uniquely incorporates spatial awareness into the channel-wise self-attention, a design innovation that maintains linear complexity, $\mathcal{O}(C^2)$, with C being the channel number. It manages to strike an impressive balance between complexity and efficiency. As shown in Tab. III, our model, carrying 139 million parameters, still situates itself within the parameter counts seen among state-of-the-art methods. More significantly, our model upholds an inference time of 125ms per image, ensuring practicality with millisecond-level response time. This efficiency does not come at the expense of performance since our model outshines competing methods in both qualitative and quantitative evaluations.

Fig. 7. Visual results of our ablation studies. A refers to replacing MPD with conventional PD, B removes the first FFN in "sandwich", C replaces SCAL with a single channel-wise self-attention design, D ablates HINT to only include channel self-attention, a single FFN, and convolutional down-sampling. E replaces our spatial branch with the basic gated mechanism from [51].

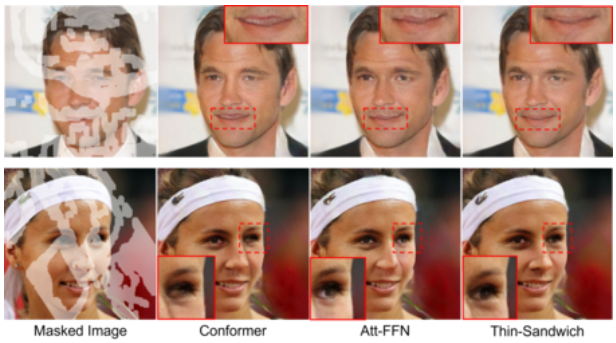

Fig. 8. Visual results of the variants of sandwich.

D. Ablation Study and Parameter Analysis

We conducted a series of ablation experiments on the CelebA-HQ dataset to evaluate the impact of each proposed component by downgrading them. All models are trained for 30,000 iterations. Our quantitative comparison results, which are presented in Tab. V, demonstrate the effectiveness of our key contributions. "U-Net w self-attention" (model D) is the variant in which we ablate HINT to only include channel self-attention [34], a single FFN, and convolutional down-sampling. We also present visual results for a more intuitive demonstration in Fig. 7.

Spatially-activated Channel Attention Layer Our proposed SCAL captures channel-wise long-range dependencies while complementing the spatial attention in an efficient manner. We suggest that introducing the spatial attention identifies "where" the important regions are. As illustrated in Fig. 7 (model C), after removing the spatial attention, the model is not confident enough to determine if an eye is missing on the left, thus generating a very blurry left eye. We substituted our spatial branch with the basic gated mechanism from [51] (model E) to evaluate our superiority. In Tab. XI, we replace the spatial branch with traditional spatial self-attention (SSA), denoted as 'w SSA', to evaluate our efficiency. However, due to the significant computational cost of SSA, we have to resize the image to 64×64 to train on a single A100. For a fair comparison, all experiments in Tab. XI are conducted on 64×64 images. We notice that the significant computational cost of SSA does not bring better performance, which is reflected in the ambiguous features with the blur texture (shown in Fig. 9).

Mask-aware Pixel-shuffle Down-sampling Our novel downsampling method based on pixel shuffling maintains a

TABLE V
ABLATION STUDIES. SETUP A REPLACES MPD WITH CONVENTIONAL PD, B REMOVES THE FIRST FFN IN "SANDWICH", C REPLACES SCAL WITH SINGLE CHANNEL-WISE SELF-ATTENTION DESIGN, D IS A HINT VARIANT WITH THE SPATIAL BRANCH REPLACED BY [51]'S GATED MECHANISM.

Setup	Model	0.01%-20%					20%-40%					40%-60%				
		PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
A	w/o MPD	34.5955	0.9649	0.4780	1.7458	0.0381	26.9292	0.8863	1.6320	4.9815	0.1084	22.5618	0.7813	3.4982	8.2196	0.1951
B	w/o Sandwich	34.7272	0.9658	0.4661	1.4687	0.0361	27.0914	0.8893	1.5796	4.7625	0.1050	22.7027	0.7853	3.4185	7.9138	0.1912
C	w/o SCAL	34.7951	0.9659	0.4624	1.7568	0.0364	27.1193	0.8895	1.5732	4.8769	0.1057	22.7206	0.7856	3.4021	8.1627	0.1925
D	U-Net w self-attention	34.0204	0.9538	0.5129	2.0152	0.0497	26.0814	0.8754	1.8547	5.1029	0.1277	21.6149	0.7679	3.6912	8.9314	0.2104
E	Full†	34.3155	0.9636	0.4891	1.3968	0.0393	26.7534	0.8837	1.6521	4.7358	0.1122	22.4632	0.7772	3.5221	7.9637	0.1999
Ours	Full	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

TABLE VI
"ATTENTION-FFN" STRUCTURE VS. "FFN-ATTENTION-FFN" STRUCTURE (SANDWICH) WITH THE SAME NUMBER OF PARAMETERS.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
SCAL-FFN	34.7272	0.9658	0.4661	1.3716	0.0361	27.0914	0.8893	1.5796	4.7174	0.1050	22.7027	0.7853	3.4185	7.8970	0.1912
Conformer	34.5125	0.9576	0.4729	1.4028	0.3914	26.9672	0.8804	1.6760	4.7597	0.1083	21.2186	0.7218	3.6829	8.9506	0.2147
Thin-Sandwich (Ours)	**34.7843**	**0.9661**	**0.4614**	**1.3697**	**0.0357**	**27.1070**	**0.8911**	**1.5763**	**4.6993**	**0.1047**	**22.7075**	**0.7872**	**3.4077**	**7.8863**	**0.1908**

TABLE VII
ABLATION STUDY OF USING 1×1 CONVOLUTION AFTER THE LAST SKIP CONNECTION.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
w 1×1 conv	34.5246	0.9646	0.4780	1.3693	0.0386	26.8984	0.8863	1.6267	4.7131	0.1101	22.4694	0.7792	3.5434	7.9647	0.1997
w/o 1×1 conv (Ours)	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

TABLE VIII
DIFFERENT KERNAL SIZE IN THE EMBEDDING LAYER.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
7×7 emb	34.6389	0.9657	0.4681	1.4034	0.0366	27.0422	0.8905	1.5803	4.9783	0.1043	22.6667	0.7865	3.3950	7.9168	0.1898
3×3 emb (Ours)	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6491**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

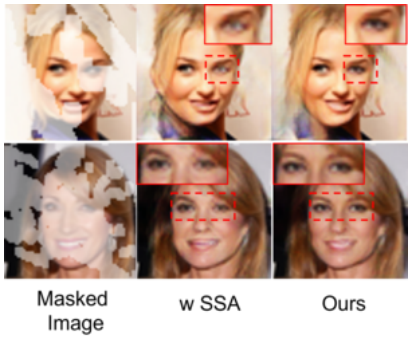

Fig. 9. Visual results of the variants of SCAL.

consistent flow of valid information within the transformer. First, to demonstrate the feasibility of pixel-shuffle down-sampling, we compare the performance of convolutional downsampling, conventional PD and the proposed MPD on the baseline. We ablate all proposed designs to build the baseline, including the "Attention-FFN" structure, single channel-wise self-attention branch, and conventional PD. As shown in Tab. IX, directly using conventional PD provides an overall improvement compared to convolutional down-sampling, but leads to a decline in LPIPS. We attribute this degradation to the incoherence of invalid information, which causes inaccurate transfer of high-level feature representations. MPD solves this problem and improves LPIPS significantly. Correspondingly, in Tab. V, the performance of HINT suffers the largest drop when we replace the MPD with conventional PD. As shown in Fig. 7 (model A), the facial attributes are severely drifting when MPD is removed.

Sandwich-shaped Transformer Block We introduce an FFN-SCAL-FFN block to effectively manage the limited flow of information. As evidenced by the results in Tab. V, removing the first FFN in the sandwich leads to a notable decrease across all four metrics. In Fig. 7 (model B), the model fails to learn a good enough feature representation of the eyeball and nose, resulting in unclear textures for the generated left eye and nose. Furthermore, to confirm that the effectiveness of the proposed Sandwich Network is not merely attributed to an increase in the number of parameters, we implemented a lightweight variant that diminishes the parameter count in both Feedforward Neural Networks (FFNs) by 50%. This thin "Sandwich" configuration possesses an equivalent number of parameters as the "Attention-FFN" architecture. Furthermore, we substituted our "FFN-SCAL-FFN" with the Conformer structure (FFN-Attention-CONV-FFN) [60] to evaluate our superiority. As shown in Fig. 8, the proposed Thin-Sandwich helps the model to learn a better feature representation of the eyeball and mouth to provide clearer texture details. Although Conformer also has a "sandwich" structure, it moves the

TABLE IX
COMPARISON OF ALTERANTIVE DESIGN OF MASK-AWARE PIXEL-SHUFFLE DOWN-SAMPLING

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
CD	34.3159	0.9641	0.4842	1.4875	0.0380	26.6809	0.8846	1.6499	4.9362	0.1088	22.2408	0.7761	3.5828	8.2493	0.1979
PD	34.5229	0.9647	0.4787	1.3729	0.0393	26.8446	0.8865	1.6328	4.8756	0.1116	22.4448	0.7795	3.5466	8.0196	0.2023
MPD (Ours)	**34.5820**	**0.9649**	**0.4733**	**1.3542**	**0.0375**	**26.9327**	**0.8867**	**1.6089**	**4.6891**	**0.1085**	**22.4769**	**0.7812**	**3.5001**	**7.8697**	**0.1972**

TABLE X
HYPER-PARAMETER TUNING ON THE WEIGHTS ASSOCIATED WITH DIFFERENT LOSSES.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
Sample A	34.6959	0.9581	0.4417	1.3143	0.0355	26.8916	0.8358	1.6470	4.8173	0.1073	22.7519	0.7850	3.4011	7.9715	0.1902
Sample B	33.5762	0.9233	0.4256	1.0158	0.0384	26.3527	0.8657	1.5419	4.9836	0.1216	22.4893	0.7754	3.4581	8.0381	0.1972
Sample C (ours)	**35.0436**	**0.9671**	**0.4489**	**1.3542**	**0.0345**	**27.2954**	**0.8924**	**1.5363**	**4.6891**	**0.1016**	**22.8473**	**0.7895**	**3.3403**	**7.8697**	**0.1867**

TABLE XI
ABLATION STUDY OF USING TRADITIONAL SPATIAL SELF-ATTENTION IN THE SCAL ON THE 64×64 RESOLUTION.

Model	0.01%-20%					20%-40%					40%-60%				
	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓	PSNR↑	SSIM↑	L1↓	FID↓	LPIPS↓
w SSA	34.4915	0.9515	0.6812	1.5641	0.0295	26.2375	0.8726	2.3536	5.0318	0.0720	21.3534	0.7726	4.3687	8.4753	**0.1312**
SCAL (Ours)	**34.5849**	**0.9538**	**0.6715**	**1.5119**	**0.0271**	**26.3195**	**0.8781**	**2.2328**	**4.9513**	**0.0707**	**21.5242**	**0.7774**	**4.2918**	**8.4518**	**0.1312**

convolutional layer that can extract local spatial feature behind the attention. Therefore, it does not embed good enough features for the following attention layer, making it difficult to generate clear texture and structure in the generated area. As shown in Table VI, the experimental results substantiate that, given an equal parameter quantity, the Sandwich module enhances the overall performance of the model.

Decision for the Last Skip Connection To harness the low-level texture and structural features derived from the encoder, we refrain from utilising 1×1 convolution for modulating the number of channels post the last skip connection. The contrast between the two strategies is enumerated in Tab. VII.

Embedding Layer In the embedding layer, we adopt a gated convolutional layer with padding to embed the input without downsampling. In contrast to prior works using 7×7 convolutional layers to project the input [10], [18], [24], a smaller kernel size (3×3) is employed in our embedding layer to obtain more fine-grained features. As illustrated in Table VIII, the smaller kernel gains better performance.

Parameter Tuning To tune HINT, we employ Optuna [70] to identify the best set of hyper-parameters in terms of different values of weights of our loss components. The top three sets of combinations are λ_i: sample A [1,1,0.5,2], sample B [1,60,1,2], sample C [1,250,0.1,0.001], as shown in Tab. X. We implement the sample C for all of the experiments.

V. CONCLUSION AND DISCUSSIONS

We propose HINT, an end-to-end Transformer for image inpainting with the proposed MPD module to ensure information remains intact and consistent throughout the encoding process. The MPD is a plug-and-play module, which is easy to adopt to the other multimedia tasks that require masking process, such as video edit, and animation edit. Our SCAL, enhanced by the proposed "sandwich" module, captures long-range dependencies while remaining spatial awareness, to boosting the capacity of representation learning in a cheap approach, which could potentially benefit multimedia tasks that are based on channel self-attention.

The proposed components contribute to each other and drive HINT to recover high-quality completed images. Experimental results demonstrate that HINT overall surpasses the current state of the art on four datasets [26]–[28], [38], with particularly notable improvements observed on facial datasets [26], [38]. Extensive qualitative evaluations demonstrate the superior image quality achieved by our framework.

As a direction of future research, HINT can be improved by employing geometric information [18], [71] by simply adding an indicator or incorporating a multi-task architecture, to get better structural consistency. Furthermore, considering the success of existing work [72], HINT can be potentially upgraded to a text-guided image inpainting system by introducing the pre-trained multi-model features to interpret the text feature into the latent space.

Furthermore, unlike existing multi-step approaches [20]–[22], as HINT is already able to recover high-quality completed images without requiring additional refinement process, a second stage of reconstruction could further enhance the quality of the results. Constrained by the current limited computing resources, we will implement another refinement network in the second step, utilizing the results from HINT as inputs and fine-tuning them in the same scale. Two networks are trained separately, thereby avoiding the large number of parameters introduced by joint training.

ACKNOWLEDGEMENTS

This research is supported in part by the EPSRC NorthFutures project (ref: EP/X031012/1).

REFERENCES

[1] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1745–1753. 1

[2] J. Wei, C. Long, H. Zou, and C. Xiao, "Shadow inpainting and removal using generative adversarial networks with slice convolutions," in *Computer Graphics Forum*, vol. 38, no. 7. Wiley Online Library, 2019, pp. 381–392. 1

[3] A. Atapour-Abarghouei and T. P. Breckon, "Dealing with missing depth: recent advances in depth image completion and estimation," *RGB-D Image analysis and processing*, pp. 15–50, 2019. 1

[4] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544. 1, 3

[5] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–14, 2017. 1, 3

[6] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 85–100. 1, 3, 6

[7] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4471–4480. 1, 3, 6, 7

[8] C. Cao and Y. Fu, "Learning a sketch tensor space for image inpainting of man-made scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 509–14 518. 1, 3

[9] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5505–5514. 1, 3, 6, 7

[10] X. Guo, H. Yang, and D. Huang, "Image inpainting via conditional texture and structure dual generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 134–14 143. 1, 3, 6, 7, 10

[11] N. Wang, J. Li, L. Zhang, and B. Du, "Musical: Multi-scale image contextual attention learning for inpainting." in *IJCAI*, 2019, pp. 3748–3754. 1, 3

[12] H. Sun, W. Li, Y. Duan, J. Zhou, and J. Lu, "Learning adaptive patch generators for mask-robust image inpainting," *IEEE Transactions on Multimedia*, 2022. 1

[13] G. Sridevi and S. Srinivas Kumar, "Image inpainting based on fractional-order nonlinear diffusion for image reconstruction," *Circuits, Systems, and Signal Processing*, vol. 38, pp. 3802–3817, 2019. 1, 2

[14] A. Atapour-Abarghouei, G. P. de La Garanderie, and T. P. Breckon, "Back to butterworth-a fourier basis for 3d surface relief hole filling within rgb-d imagery," in *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 2016, pp. 2813–2818. 1, 2

[15] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009. 1, 2

[16] R. Zhang, W. Quan, Y. Zhang, J. Wang, and D.-M. Yan, "W-net: Structure and texture interaction for image inpainting," *IEEE Transactions on Multimedia*, 2022. 1, 6, 7

[17] Y. Zhang, X. Zhang, C. Shi, X. Wu, X. Li, J. Peng, K. Cao, J. Lv, and J. Zhou, "Pluralistic face inpainting with transformation of attribute information," *IEEE Transactions on Multimedia*, 2022. 1

[18] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," *arXiv preprint arXiv:1901.00212*, 2019. 1, 3, 5, 7, 10

[19] S. Uddin and Y. J. Jung, "Global and local attention-based free-form image inpainting," *Sensors*, vol. 20, no. 11, p. 3204, 2020. 1, 2

[20] C. Zheng, T.-J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 512–11 522. 1, 2, 3, 4, 5, 10

[21] Z. Wan, J. Zhang, D. Chen, and J. Liao, "High-fidelity pluralistic image completion with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4692–4701. 1, 2, 3, 4, 5, 6, 7, 8, 10

[22] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 758–10 768. 1, 2, 3, 4, 5, 6, 7, 8, 10

[23] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159. 1, 7

[24] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1869–1878. 1, 5, 6, 7, 10

[25] Q. Guo, X. Li, F. Juefei-Xu, H. Yu, Y. Liu, and S. Wang, "Jpgnet: Joint predictive filtering and generative network for image inpainting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 386–394. 1, 7

[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017. 1, 2, 5, 6, 10

[27] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017. 1, 2, 5, 6, 10

[28] T. Yu, S. Zhang, C. Lin, S. You, J. Wu, J. Zhang, X. Ding, and H. An, "Dunhuang grottoes painting dataset and benchmark," *arXiv preprint arXiv:1907.04589*, 2019. 1, 2, 5, 7, 8, 10

[29] G. Zhao, J. Wang, Z. Zhang *et al.*, "Random shifting for cnn: a solution to reduce information loss in down-sampling layers." in *IJCAI*, 2017, pp. 3476–3482. 2

[30] S. A Sharif, R. A. Naqvi, and M. Biswas, "Beyond joint demosaicking and denoising: An image processing pipeline for a pixel-bin image sensor," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 233–242. 2

[31] Z. Yue, J. Xie, Q. Zhao, and D. Meng, "Semi-supervised video deraining with dynamical rain generator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 642–652. 2

[32] L. Wang, Y. Wang, Z. Lin, J. Yang, W. An, and Y. Guo, "Learning a single network for scale-arbitrary super-resolution," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4801–4810. 2

[33] Y. Zhou, J. Jiao, H. Huang, Y. Wang, J. Wang, H. Shi, and T. Huang, "When awgn-based denoiser meets real noises," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 13 074–13 081. 2

[34] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739. 2, 3, 4, 5, 8

[35] M. Li, Y. Fu, and Y. Zhang, "Spatial-spectral transformer for hyperspectral image denoising," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1368–1376. 2

[36] S.-I. Jang, T. Pan, Y. Li, P. Heidari, J. Chen, Q. Li, and K. Gong, "Spach transformer: spatial and channel-wise transformer based on local and global self-attentions for pet image denoising," *IEEE Transactions on Medical Imaging*, 2023. 2

[37] L. Wang, M. Cao, Y. Zhong, and X. Yuan, "Spatial-temporal transformer for video snapshot compressive imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[38] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738. 2, 5, 7, 8, 10

[39] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *arXiv e-prints*, pp. arXiv–1406, 2014. 3

[40] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, "Recurrent feature reasoning for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7760–7768. 3, 7

[41] J. Peng, D. Liu, S. Xu, and H. Li, "Generating diverse structure for image inpainting with hierarchical vq-vae," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 775–10 784. 3

[42] H. Wu, J. Zhou, and Y. Li, "Deep generative model for image inpainting with local binary pattern learning and spatial attention," *IEEE Transactions on Multimedia*, vol. 24, pp. 4016–4027, 2021. 3

[43] T. Yu, Z. Guo, X. Jin, S. Wu, Z. Chen, W. Li, Z. Zhang, and S. Liu, "Region normalization for image inpainting," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 733–12 740. 3

[44] H. Zheng, Z. Lin, J. Lu, S. Cohen, E. Shechtman, C. Barnes, J. Zhang, N. Xu, S. Amirghodsi, and J. Luo, "Image inpainting with cascaded modulation gan and object-aware training," in *European Conference on Computer Vision*. Springer, 2022, pp. 277–296. 3

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017. 3

[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022. 3, 4

[47] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[48] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, "Diverse image inpainting with bidirectional and autoregressive transformers," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 69–78. 3

[49] Y. Zhang, Y. Liu, R. Hu, Q. Wu, and J. Zhang, "Mutual dual-task generator with adaptive attention fusion for image inpainting," *IEEE Transactions on Multimedia*, 2023. 3

[50] Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang, "Learning contextual transformer network for image inpainting," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2529–2538. 3

[51] ——, "T-former: An efficient transformer for image inpainting," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 6559–6568. 3, 8, 9

[52] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016. 3

[53] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in Neural Information Processing Systems*, vol. 34, pp. 30 392–30 400, 2021. 3

[54] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883. 3, 4

[55] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258. 4, 5

[56] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154. 4

[57] K. Fukushima, "Cognitron: A self-organizing multilayered neural network," *Biological cybernetics*, vol. 20, no. 3-4, pp. 121–136, 1975. 4

[58] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *Proc. Interspeech 2019*, pp. 2613–2617, 2019. 5

[59] D. S. Park, Y. Zhang, C.-C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6879–6883. 5

[60] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020. 5, 9

[61] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 10 684–10 695. 6, 7, 8

[62] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "Wavefill: A wavelet-based generation network for image inpainting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 114–14 123. 6, 7, 8

[63] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 11 461–11 471. 6, 7, 8

[64] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 181–190. 7

[65] D. P. Kingma, J. A. Ba, and J. Adam, "A method for stochastic optimization. arxiv 2014," *arXiv preprint arXiv:1412.6980*, vol. 106, 2020. 6

[66] Y. Zeng, J. Fu, H. Chao, and B. Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1486–1494. 6

[67] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 6

[68] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595. 6

[69] Z. Chang, G. A. Koulieris, and H. P. H. Shum, "On the design fundamentals of diffusion models: A survey," *arXiv*, 2023. [Online]. Available: http://arxiv.org/abs/2306.04542 8

[70] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631. 10

[71] S. Chen, A. Atapour-Abarghouei, J. Kerby, E. S. L. Ho, D. C. G. Sainsbury, S. Butterworth, and H. P. H. Shum, "A feasibility study on image inpainting for non-cleft lip generation from patients with cleft lip," in *Proceedings of the 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics*, ser. BHI '22. IEEE, 9 2022, pp. 1–4. 10

[72] M. Ni, X. Li, and W. Zuo, "Nuwa-lip: Language-guided image inpainting with defect-free vqgan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 183–14 192. 10

Shuang Chen received his B.S. degree in Automation from Shandong University of Technology in 2017, and MSc. degree in Computer Vision, Machine Learning and Robotics from University of Surrey in 2020. He is currently pursuing his PhD degree at the Department of Computer Science at Durham University in the UK. His research interests span across Computer Vision, Machine Learning, Image Processing and Image Inpainting.

Amir Atapour-Abarghouei is an Assistant Professor at the Department of Computer Science at Durham University. Prior to his current post, Amir held a lectureship position at Newcastle University in the UK. He received his PhD from Durham University. His primary research interests span across Computer Vision, Machine Learning and Natural Language Processing, Anomaly Detection, Brain-Computer Interface and Graph Analysis. His work includes the generalised high-impact GANomaly anomaly detection approach, which is now a part of Intel's AI product line and used as the underlying method for anomaly detection in numerous international patents. Amir has co-organised the CVPR-NAS workshop as well as workshops at BigData (BDA4CID and BDA4HM).

Hubert P. H. Shum (Senior Member, IEEE) is an Associate Professor in Visual Computing and the Deputy Director of Research of the Department of Computer Science at Durham University, specialised in Spatio-Temporal Modelling and Responsible AI. He is also a Co-Founder of the Responsible Space Innovation Centre. Before this, he was an Associate Professor at Northumbria University and a Post-doctoral Researcher at RIKEN Japan. He received his PhD degree from the University of Edinburgh. He chaired conferences such as Pacific Graphics, BMVC and SCA, and has authored over 150 research publications.