CoReaP : Collaborative Reconstruction with assistive Priors

Chanhee Lee* Mingu Kang* Gahyun Kim* Yeeun Hwang*
Sungkyunkwan University, South Korea
{leechanhye, gms5560, soda3042, yeeun89}@g.skku.edu

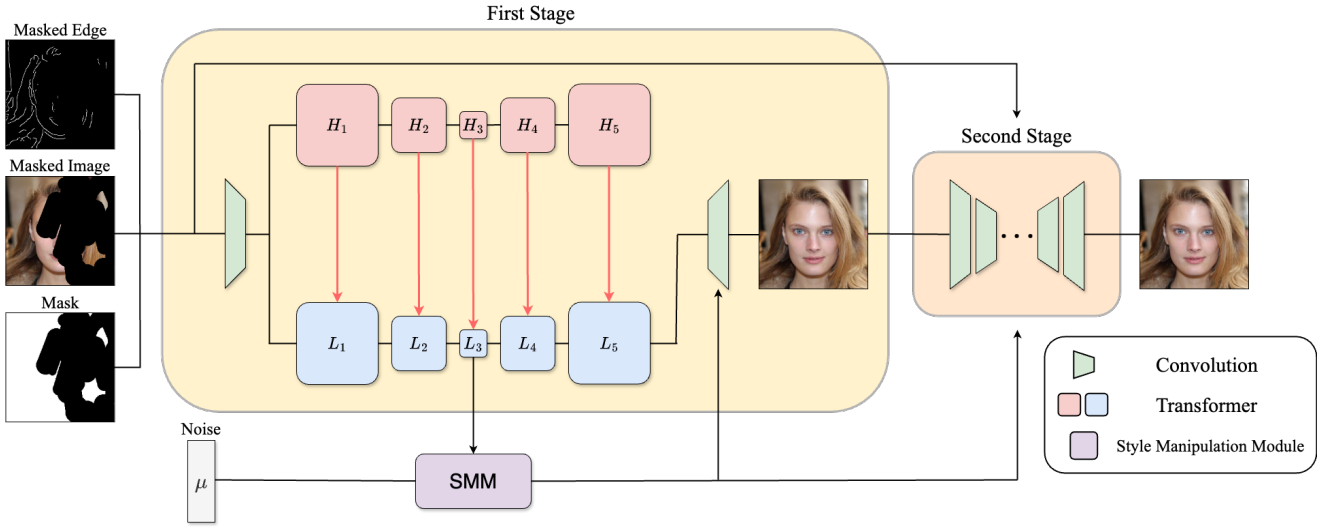


Figure 1. An overview of the CoReaP architecture. Here, H and L denote transformer blocks in the high-frequency and low-frequency paths, respectively. CoReaP first completes the masked regions in the First Stage and subsequently refines fine details in the Second Stage through a reconstruction process with the style manipulation module [17].

Abstract

Image inpainting aims to restore missing regions in images by generating semantically plausible and visually coherent content. A key challenge in this task is effectively leveraging unmasked regions to guide the reconstruction process. While recent advancements have primarily focused on user-guided content generation, we propose a novel approach that fundamentally enhances inpainting capability. Specifically, we introduce ***CoReaP*** *: Collaborative Reconstruction with Assistive Priors, a two-path learning framework that separately processes high-frequency and low-frequency features to achieve more structured and detailed reconstruction. High-frequency features act as priors to guide the low-frequency path, while the low-frequency outputs further refine the inpainting process. To improve feature aggregation and receptive field expansion, we integrate deformable convolution into the transformer-based tokenization process. Our method,* ***CoReaP****, introduces an innovative two-path architecture where the high-frequency path serves as assistive priors, enhancing the low-frequency path's ability to reconstruct complex structures and fine details. This synergistic interaction offers new insights for related tasks that require the integration of multi-frequency information for more coherent and visually plausible outcomes. Our source code is available at https://github.com/rkdrn79/CoReaP.*

*Equal contribution.

1. Introduction

Image inpainting aims to reconstruct masked regions in images by generating visually coherent and contextually plausible content. Achieving plausible mask completion relies on effectively aggregating information from unmasked regions, which is essential for successful image reconstruction. Image inpainting is a critical task in computer vision serving as a benchmark for assessing the capability of new architectures in understanding image context, texture, and structural composition. While recent developments in image inpainting have predominantly emphasized user-guided

content generation [21, 31, 37, 38, 46], our approach diverges from this trajectory by prioritizing the fundamental improvement of the model's inpainting capability.

Successful inpainting requires the ability to learn both (1) plausible semantic structures and (2) fine-grained details [15], including texture, lines, and edges. To the best of our knowledge, these fine-grained details are preceded by plausible semantic structures; that is, filling the masked region with plausible content is followed by detail refinement [15, 17, 32]. This perspective suggests that low-frequency features act as priors for high-frequency details, providing essential guidance for the model in refining the masked regions with appropriate fine details [26]. This highlights the role of priors in simplifying the inpainting process. Conversely, attempting inpainting without priors significantly increases the complexity of the task. Furthermore, learning low-frequency information presents a greater challenge than high-frequency feature learning, as it requires capturing global interactions, whereas high-frequency learning primarily relies on local interactions [1].

In this paper, we introduce a two-path learning framework that separately processes high-frequency and low-frequency features. High-frequency features act as priors to guide the low-frequency path, while the generated images from the low-frequency path further serve as priors for the refinement stage. This approach enables the generation of more visually coherent images compared to previous methods. The low-frequency path produces an enhanced image, which is subsequently refined in the Second-stage, where both low-frequency and high-frequency features contribute to further improving plausibility. To implement this framework, we integrate deformable convolution, taking into account the distinct characteristics of low-frequency and high-frequency learning.

Our contributions are summarized as:

- We incorporate DCN into the tokenization process of the transformer block, enabling effective aggregation of unmasked region information and facilitating rapid receptive field expansion.
- We explicitly enhance high-frequency learning by introducing a dedicated path optimized with high-frequency loss. Furthermore, the extracted high-frequency features contribute to the low-frequency path, facilitating the generation of more coherent and realistic content within the masked regions.
- CoReaP introduces a novel frequency-guided architecture that leverages high-frequency priors to assist the low-frequency path, enhancing structural coherence and detail reconstruction.

Our approach presents a novel architecture that leverages frequency-guided processing, incorporating DCN within the transformer's tokenization stage. This design allows the model to efficiently capture information from unmasked regions while rapidly expanding its receptive field for enhanced contextual understanding. By explicitly optimizing the high-frequency path with a dedicated loss, the extracted high-frequency features serve as assistive priors for the low-frequency path. This interaction enhances structural coherence, improves detail reconstruction, and facilitates the generation of visually plausible content within masked regions.

2. Related Works

Recent advances in image inpainting have leveraged diverse deep learning architectures to address the dual challenges of contextual coherence and detail preservation. We review three critical directions: (1) transformer-based approaches for global context modeling, (2) methods addressing long-range dependencies in large missing regions, and (3) techniques incorporating frequency-aware processing to enhance detail reconstruction.

2.1. Transformer-Based Approaches in Image Inpainting

The adoption of transformers in image inpainting has grown significantly since their success in natural language processing [28] and vision tasks [7]. Unlike convolutional networks constrained by local receptive fields, transformers utilize self-attention mechanisms to model long-range dependencies - a critical capability for coherent reconstruction of missing regions [17].

Early explorations like the Image Processing Transformer (IPT) [3] demonstrated transformers' potential for low-level vision tasks. Subsequent work improved efficiency and quality through innovations like the Spatial Diffusion Model (SDM) [18] with iterative probabilistic refinement, and the T-former [5] introducing resolution-linear attention. Recent architectures hybridize transformer strengths with inductive biases from CNNs: The Inpainting Transformer (ITrans) [43] combines self-attention with convolution, while TransInpaint [24] implements context-adaptive attention conditioned on mask geometry. Mask-Aware Transformer (MAT) [17] advances this direction with token validity guidance, achieving high-fidelity results on challenging masks through focused non-local aggregation.

Despite progress, existing methods treat inpainting as a monolithic task without explicit frequency decomposition. This often results in blurred textures and compromised edges due to inadequate high-frequency handling [20]. Our work addresses this limitation through dual-path processing of frequency components, enabling targeted detail reconstruction while maintaining global coherence.

2.2. Long-Range Dependency Modeling

Addressing large masks requires modeling relationships between distant valid tokens - a weakness of CNNs due

to gradual receptive field expansion [12]. While attention mechanisms theoretically solve this through global interactions [28], practical implementations face challenges beyond quadratic complexity.

In large mask scenarios, sparse valid tokens lead attention to over-weight limited information pockets [17]. This "token starvation"problem constrains model capacity, as evidenced by performance drops on masks exceeding 50% coverage [39]. Recent solutions like deformable attention [45] and token grouping [2] aim to mitigate this, but retain computational overheads. Our architecture introduces dynamic token selection that adapts to mask topology, balancing efficiency with comprehensive context utilization.

2.3. Inpainting with High-Frequency Guidance

Frequency-aware processing, inspired by successes in video analysis [8], shows growing promise for detail preservation. The Zoom-In Transformer (ZITS) [6] pioneered edge-aware inpainting through wavelet decomposition, while HIFILL [30] demonstrated frequency-specific feature modulation. However, these approaches process components separately without cross-frequency guidance.

Our **early fusion** strategy draws inspiration from SlowFast networks [8], where high and low frame rate pathways exchange temporal-spatial information. Reinterpreted for images, we maintain distinct processing streams for different frequency bands while enabling progressive integration—preserving high-frequency details through direct feature injection rather than late-stage concatenation [30].

3. Methods

Our method is divided into three sections. The first section is **tokenization with deformable convolution**. The second section is **dividing the existing path into two paths**; high-frequency and low-frequency. And the third section is **early fusion**, which fuses high-frequency features into low-frequency path in each stage using deformable convolution to handle misalignment, caused by different tokenization applied.

3.1. Overall Architecture

As illustrated in Fig. 1, the proposed **CoReaP** architecture is divided into two stages: the First Stage and the Second Stage. The First Stage comprises a convolutional head, a transformer body, and a convolutional tail, which are responsible for initial feature extraction and representation learning. The Second Stage incorporates a style manipulation module, which further refines and enhances the processed features by reconstruction [17].

First stage, Downsampling masked images in the **convolutional head** before they are fed into the transformer block is not only enhances computational and memory efficiency but also semantic representation. That is, processing image in the early stage with inductive bias is vital for optimization [36] and better performance [17, 34]. A transformer with five stages having different resolutions has two paths; **the high-frequency path and the low-frequency path**. The high frequency path focuses on the generation of high frequency elements. In contrast, the low-frequency path takes on a duty of generating plausible contents considering valid tokens that have richer information while dismissing invalid tokens (their values are nearly zero [17]). During each stage of the transformer body in the First-stage, the high-frequency features are integrated into the low-frequency path. It gives auxiliary information to the low-frequency path which lacks high-frequency information.

Second stage, we used **U-Net shaped Conv layer** to refine the details of the images generated from the low-frequency path, and a style manipulation module to generate more diverse inpainted images [17].

3.2. Deformable Convolution Tokenization

While recent advances in image inpainting have demonstrated promising results, existing approaches face critical limitations when handling large masked regions. CNN-based methods suffer from (1) inefficient expansion of receptive fields [26], as their rigid convolutional kernels struggle to adaptively capture long-range dependencies. Conversely, attention-based architectures address this issue but introduce two new bottlenecks: (2) quadratic computational complexity and (3) over-concentration on sparse valid tokens, where standard attention mechanisms disproportionately amplify isolated valid regions while underutilizing the broader contextual diversity in unmasked areas. These limitations hinder the model's ability to generate coherent and realistic inpainting results, especially for large masked regions To address challenges (1) and (3), we propose Deformable Convolution Tokenization (**DCT**), a feature transformation module within the transformer body (Fig. 2). Unlike conventional tokenization with fixed-grid sampling or just flattening, DCT leverages Deformable Convolution Networks (DCN)—originally designed for geometric-invariant object detection [4]—to decouple the sampling of valid tokens from spatial rigidity. By learning dynamic sampling offsets, DCT achieves:

- Rapid Receptive Field Expansion: Adaptive kernel offsets enable the model to gather features from distant valid regions within a single layer, bypassing the gradual receptive field growth of standard CNNs.
- Diversified Token Selection: Offset-driven sampling strategically aggregates tokens from spatially dispersed yet semantically critical unmasked areas, mitigating attention's tendency to overfit to local valid patches.

This repurposing of DCN diverges from its original geometric invariance objective, instead harnessing its input-

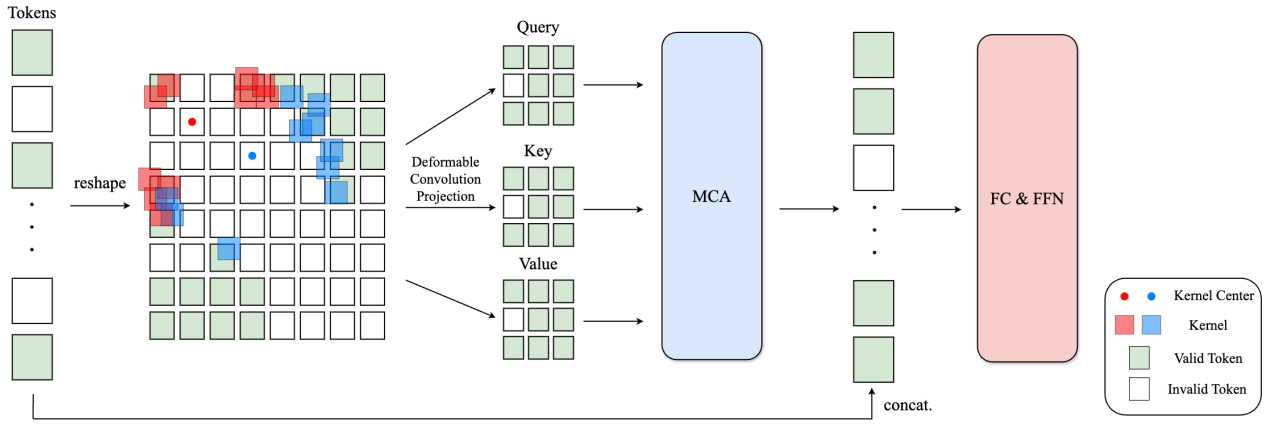


Figure 2. The transformer block incorporating Deformable Convolution Tokenization (DCT) follows a structured pipeline. Initially, the input is projected into Query, Key, and Value representations using DCT, followed by Multi-head Contextual Attention (MCA) [17]. The output of MCA is then concatenated with the original input and fused through a fully connected (FC) layer. Finally, the processed features are passed into a Feed-Forward Network (FFN), maintaining the standard structure of the Transformer [28].

dependent adaptability to dynamically reconfigure token sampling paths. This capability critical for large-mask inpainting, where valid context is sparse and irregularly distributed. DCT builds upon convolution tokenization in CvT [35] while integrating the adaptability of DCN. Notably, it transforms non-overlapping tokenization into an overlapping tokenization scheme, enhancing feature continuity and spatial information retention. It employs deformable convolution projection to generate Query, Key, and Value within the transformer body, as opposed to relying solely on linear projection or convolution projection with the same stride and kernel size for token generation from a $B \times L \times C$-shaped, flattened feature map:

$$Q, K, V = \text{FC}((\text{Flatten}(X^H_{k,i-1})) \ , \tag{1}$$

$$Q, K, V = \text{FC}(\text{Flatten}(\text{DCN}(X^L_{k,i-1}))) \ , \tag{2}$$

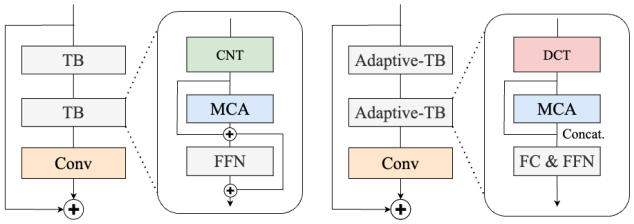
where $X^H_{k,i}$ denotes the feature representation at the i-th layer of the high-frequency path, while $X^L_{k,i}$ denotes that of the low-frequency path in the k-th stage. Each of Q, K, and V corresponds to Query, Key, and Value, respectively, and is processed at independent modules. DCT facilitates the retention of valid tokens for the majority of tokens while filtering out redundant ones within the attention module, enabling the model to effectively select relevant information from a rich and diverse token space. In our implementation, DCNv3 [33] is utilized in place of DCNv1 [4] for improved performance and efficiency.

3.3. Dividing a Path into Two Paths

The First Stage is divided into two separate paths to extract high-frequency and low-frequency features, which are leveraged to generate more visually plausible images before detail refinement in the Second Stage. Residual connections [10] have emerged as an essential component in deep learning, facilitating stable and easier training by providing additional pathways that alleviate the vanishing gradient problem [10]. This approach allows models to learn residuals at each layer by formulating the target function as the sum of the residual function and the identity: $\mathcal{H}(x) := \mathcal{F}(x) + x$. Here, $\mathcal{H}(x)$ denotes the target function, $\mathcal{F}(x)$ represents the residual function, and x corresponds to the identity mapping, which is equivalent to the input. The residual learning mechanism can be interpreted as facilitating the model's ability to capture high-frequency bases [17].

Given that our architecture focuses on detail refinement in the Second Stage, ensuring the generation of plausible content in the masked regions during the First Stage is crucial. However, requiring the model to complete a masked region on a blank canvas often leads to ambiguity. To alleviate ambiguity, we hypothesize that the model benefits from priors, such as sketches composed of edges and lines, to generate more plausible content. Therefore, we partition the existing path in the First Stage into two distinct pathways: high-frequency and low-frequency. The high-frequency features serve as priors to guide the low-frequency path, enhancing the reconstruction of masked regions with more plausible content, solving "inter-frequency conflicts"[40]. The integration of high-frequency priors into the low-frequency path is further detailed in Section 3.4.

Building on the observation that learning patterns can be modulated through the inclusion or exclusion of residual connections [17], we designed the high-frequency and low-frequency paths with distinct structural configurations. Within the high-frequency path, residual connections are incorporated to implicitly guide the model's learning process as shown in Fig. 3a. Following the transformer body, a loss



(a) High Frequency (b) Low Frequency

Figure 3. Discrepancy between high & low frequency path. CNT denotes CNN Tokenization, whereas DCT represents DCN Tokenization.

function (e.g., focal loss [23]) is applied to explicitly facilitate the learning of high-frequency bases. Conversely, the low-frequency path is designed to capture low-frequency bases. To facilitate this, residual connections are removed (Fig. 3b). Each path's output after MCA is expressed as:

$$\hat{X}^H_{k,i} = \text{MCA}(X^H_{k,i}) + X^H_{k,i} \ , \tag{3}$$

$$\hat{X}^L_{k,i} = \text{FC}([\text{MCA}(X^L_{k,i}) \ ; \ X^L_{k,i}]) \ . \tag{4}$$

3.4. Early Fusion

To effectively integrate the high-frequency features extracted from the high-frequency path with the low-frequency features, an appropriate fusion strategy must be devised. As illustrated in Fig. 3, DCT is applied exclusively to the low-frequency path, while the high-frequency path is applied the standard CNN tokenization. To enhance the model's ability to capture local image details for learning high-frequency bases [1], CNN-based tokenization is employed in the high-frequency path with relatively small kernels, preserving fine-grained spatial information.

However, the inherent discrepancy between tokenization methods introduces challenges in accurately fusing high- and low-frequency features. In other words, at the same stage, high-frequency features possess a smaller receptive field than low-frequency features due to the adaptive dilation effect of DCN's offset in the tokenization layer, which facilitates rapid receptive field expansion. This incurs misalignment between the two features. Accordingly, DCN is incorporated into early fusion to mitigate misalignment, consistent with its intended original design objective. The early fusion mechanism incorporating DCN is defined as follows

$$X^L_k = \text{Conv}([\dot{X}^L_k \ ; \ \text{DCN}(\dot{X}^H_k)]) \ , \tag{5}$$

where $\dot{X}k^L$ and $\dot{X}k^H$ denote the output features of the k-th stage following the completion of all layer operations within that stage. The updated X^L_k is propagated to the next stage within the low-frequency path, whereas $\dot{X}^H_k$ advances to the next stage without modification.

3.5. Loss Function

We use adversarial loss [9], pereceptual loss [14], and R1 regularization [19, 22] to enhance quality and diversity of inpainted images, following the training strategies of state-of-the-art models [2, 17]. Adversarial loss is defined as:

$$\mathcal{L}_G = -\mathbb{E}_{\hat{I}}[\log D(\hat{I})] \ , \tag{6}$$

$$\mathcal{L}_D = -\mathbb{E}_I[\log D(I)] - \mathbb{E}_{\hat{I}}[\log(1 - D(\hat{I}))] \ , \tag{7}$$

where G and D denote the generator and the discriminator, respectively, while $\hat{I}$ and I correspond to the generated images and the ground truth.

Perceptual loss evaluates the similarity between the output features of the generated images and those of the ground truth using a pretrained CNN model, as defined below:

$$\mathcal{L}_P = \sum_i \lambda^P_i ||\phi(\hat{I}) - \phi(I)||_1 \ , \tag{8}$$

where $\phi(\cdot)$ denoted the output features of a VGG-19 [25] network.

Focal Loss Residual connections are integrated into the high-frequency path to support implicit high-frequency learning. Meanwhile, a dedicated loss function is applied to explicitly enhance the model's ability to capture high-frequency features. Specifically, focal loss [23] is employed for the high-frequency path, as ground truth edge images predominantly contain black pixels, with edge components being relatively sparse.

$$\mathcal{L}_F = \sum_i -\alpha_i (1 - p_i)^\gamma \log(p_i) \ . \tag{9}$$

4. Experiments

In this section, we assess the effectiveness of CoReaP in the mask inpainting task by comparing its performance against existing state-of-the-art methods. We conduct both quantitative and qualitative evaluations to comprehensively analyze the model's capability in reconstructing missing regions while preserving perceptual quality and semantic consistency.

4.1. Dataset and Evalutation

Our experiments are conducted on the CelebA-HQ [13] dataset at a resolution of 256 × 256. Specifically, we utilize 28,000 images from the training set and 2,000 images from the validation set to train and evaluate our model. The dataset provides high-quality facial images, making it well-suited for assessing the effectiveness of inpainting models in generating realistic and coherent reconstructions. To

standardize the input, all images are resized to 256 × 256, ensuring consistency across the dataset. Additionally, we apply random masking to simulate missing regions, creating diverse inpainting scenarios. The masks vary in shape and size, challenging the model to reconstruct plausible structures while maintaining perceptual fidelity.

To evaluate the effectiveness of our inpainting model, we employ three widely used metrics: *Fréchet Inception Distance* (FID) [11], *Perceptual Image Deviation Score* (P-IDS) [42], and *Uncertainty-aware Image Deviation Score* (U-IDS) [16]. These metrics collectively assess the perceptual realism, fidelity, and consistency of the generated images. FID [11] is used to measure the similarity between the distribution of generated and real images. By computing the Fréchet distance between feature embeddings extracted from an Inception network, FID quantifies how closely the generated images resemble real ones, with lower values indicating better quality. While FID evaluates overall distributional similarity, P-IDS provides a more localized assessment by directly comparing inpainted images with their corresponding ground truth. It leverages deep feature representations to measure perceptual differences, capturing fine-grained details that pixel-wise metrics may overlook [42]. A lower P-IDS score suggests that the inpainted image better preserves the structure and content of the original. Since inpainting models often exhibit variations in their outputs depending on the masked regions, U-IDS extends P-IDS by incorporating uncertainty estimation [16]. By evaluating the consistency of multiple inpainted results for the same input, U-IDS ensures that the model generates stable and reliable reconstructions. A lower U-IDS score indicates higher robustness in handling diverse missing regions. Together, these metrics provide a comprehensive evaluation framework, enabling a thorough assessment of both the realism and reliability of our inpainting model.

4.2. Quantitative results

We present the numerical evaluation of our method in comparison with prior approaches. Our model demonstrates superior performance across multiple metrics, highlighting its ability to produce high-quality inpainted images. The detailed results are provided in Table 1.

4.3. Qualitative Results

To further illustrate the effectiveness of CoReaP, we visualize the inpainting results on representative samples. Figure **??** showcases the reconstructed images alongside the corresponding ground truth and masked inputs. Our approach generates perceptually realistic outputs, preserving fine-grained details and maintaining structural coherence better than competing methods.

Method	Small Mask			Large Mask		
	FID↓	P-IDS↑	U-IDS↑	FID↓	P-IDS↑	U-IDS↑
CoReaP(Ours)	-	-	-	-	-	-
MAT [17]	**2.94**	**20.88**	**32.01**	**5.16**	**13.90**	**25.13**
LaMa [27]	3.98	8.82	22.57	8.75	2.34	8.77
ICT [29]	5.24	4.51	17.39	10.92	0.90	5.23
MADF [44]	10.43	6.25	14.62	23.59	0.50	1.44
AOT GAN [41]	9.64	5.61	14.62	22.91	0.47	1.65
DeepFill v2 [39]	5.69	6.62	16.82	13.23	0.84	2.62
EdgeConnect [20]	5.24	5.61	15.65	12.16	0.84	2.31

Table 1. Quantitative results on CelebA-HQ at 256×256 size. The results of P-IDS and U-IDS are shown in percentage (%).

5. Conclusion

In this paper, we present **CoReaP**, a novel two-path framework for image inpainting that explicitly leverages frequency-guided learning to enhance structural coherence and fine-grained detail reconstruction. By decoupling the learning of low-frequency and high-frequency features, our approach addresses the inherent challenges of simultaneously capturing global context and local details. The low-frequency path, guided by high-frequency priors, generates semantically plausible structures, while the high-frequency path refines textures and edges using dedicated optimization. The integration of deformable convolution (DCN) within the transformer's tokenization stage further enables efficient aggregation of unmasked region information and rapid expansion of the receptive field, critical for contextual understanding.

CoReaP is expected to outperform existing methods in generating visually coherent and contextually consistent inpainted content. The explicit separation of frequency learning pathways, coupled with bidirectional feature guidance, ensures that structural priors inform detail refinement and vice versa, reducing artifacts and improving realism. The success of this framework underscores the importance of frequency-aware processing in complex vision tasks and provides a blueprint for future architectures seeking to balance global and local feature learning.

Future work will explore extending this frequency-guided paradigm to other image restoration tasks, such as super-resolution and denoising, where hierarchical feature interaction is equally critical. Additionally, investigating dynamic mechanisms to adaptively weight low-frequency and high-frequency contributions based on mask characteristics could further enhance inpainting performance. Our approach reaffirms the value of fundamental model improvements in advancing computer vision, paving the way for more robust and versatile image synthesis systems.

References

[1] Genggeng Chen, Kexin Dai, Kangzhen Yang, Tao Hu, Xiangyu Chen, Yongqing Yang, Wei Dong, Peng Wu, Yanning Zhang, and Qingsen Yan. Bracketing image restoration and enhancement with high-low frequency decomposition. *arXiv preprint arXiv:2404.13537*, 2024. 2, 5

[2] Haiwei Chen and Yajie Zhao. Don't look into the dark: Latent codes for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7591–7600, 2024. 3, 5

[3] Hanting Chen et al. Pre-trained image processing transformer. *CVPR*, 2021. 2

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3, 4

[5] Ye Deng, Siqi Hui, Sanping Zhou, Deyu Meng, and Jinjun Wang. T-former: An efficient transformer for image inpainting. *arXiv preprint arXiv:2305.07239*, 2023. 2

[6] Yin Dong et al. Zits: Zoom-in transformer for high-resolution image inpainting. *CVPR Workshop*, 2022. 3

[7] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2020. 2

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 3

[13] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[14] Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019. 5

[15] Soo Ye Kim, Kfir Aberman, Nori Kanazawa, Rahul Garg, Neal Wadhwa, Huiwen Chang, Nikhil Karnad, Munchurl Kim, and Orly Liba. Zoom-to-inpaint: Image inpainting with high-frequency details. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 477–487, 2022. 2

[16] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 6

[17] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022. 1, 2, 3, 4, 5, 6

[18] Wenbo Li, Xin Yu, Kun Zhou, Yibing Song, Zhe Lin, and Jiaya Jia. Image inpainting via iteratively decoupled probabilistic modeling. *arXiv preprint arXiv:2212.02963*, 2022. 2

[19] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International conference on machine learning*, pages 3481–3490. PMLR, 2018. 5

[20] K Nazeri. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2, 6

[21] Minheng Ni, Xiaoming Li, and Wangmeng Zuo. Nuwa-lip: language-guided image inpainting with defect-free vqgan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14192, 2023. 2

[22] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 5

[23] T-YLPG Ross and GKHP Dollár. Focal loss for dense object detection. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2980–2988, 2017. 5

[24] Pourya Shamsolmoali, Masoumeh Zareapoor, and Eric Granger. Transinpaint: Transformer-based image inpainting with context adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 849–857, 2023. 2

[25] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[26] Lin Sun, Bin Jiang, Chao Yang, Jiawu Dai, and Weiyuan Zeng. Repgan: image inpainting via residual partial connection and mask discriminator. *International Journal of Machine Learning and Cybernetics*, 14(9):3193–3203, 2023. 2, 3

[27] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022. 6

[28] Ashish Vaswani et al. Attention is all you need. *NeurIPS*, 2017. 2, 3, 4

[29] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transform-

ers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4692–4701, 2021. 6

[30] Ziyu Wan et al. Hifill: Image inpainting via hierarchical feature-aware learning. *IJCAI*, 2021. 3

[31] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18359–18369, 2023. 2

[32] Wentao Wang, Li Niu, Jianfu Zhang, Xue Yang, and Liqing Zhang. Dual-path image inpainting with auxiliary gan inversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11421–11430, 2022. 2

[33] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023. 4

[34] W Weng and X Zhu INet. Convolutional networks for biomedical image segmentation., 2021, 9. *DOI: https://doi.org/10.1109/ACCESS*, pages 16591–16603, 2021. 3

[35] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021. 4

[36] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in neural information processing systems*, 34:30392–30400, 2021. 3

[37] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023. 2

[38] Jianjin Xu, Saman Motamed, Praneetha Vaddamanu, Chen Henry Wu, Christian Haene, Jean-Charles Bazin, and Fernando De la Torre. Personalized face inpainting with diffusion models by parallel visual attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5432–5442, 2024. 2

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 3, 6

[40] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14114–14123, 2021. 4

[41] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 29(7):3266–3280, 2022. 6

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[43] Yifan Zhang, Zhen Li, Lei Zhang, and Bo Zhang. Itrans: Generative image inpainting with transformers. *Multimedia Systems*, 2023. 2

[44] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, Errui Ding, and Zhaoxiang Zhang. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021. 6

[45] Xizhou Zhu et al. Deformable transformers for end-to-end object detection. *ICLR*, 2021. 3

[46] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *European Conference on Computer Vision*, pages 195–211. Springer, 2024. 2