Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 9
AIDML 2025

Advances in Image Inpainting: Global Context Modeling via Transformers and Diffusion Models

Jiaoyang Li *

School of Data Science, Chinese University of Hong Kong, Shenzhen, China

* Corresponding Author Email: 122040058@link.cuhk.edu.cn

Abstract. Image inpainting, a critical task in computer vision, has significantly benefited from the rapid development of deep learning techniques, particularly Transformers and Diffusion Models. Traditional methods relying on texture matching and PDE-based diffusion strategies demonstrate limited effectiveness in complex or extensive damaged regions. Recent advancements employing Transformer architectures effectively exploit global context via self-attention mechanisms, ensuring structural coherence in large missing areas. Hybrid models integrating transformers and convolutional networks, such as MAT, further enhance performance by combining global semantic understanding and local detail restoration. Meanwhile, diffusion Models, through iterative denoising steps, offer substantial improvements in realism and texture fidelity, outperforming previous methods in generating high-quality, diverse inpainting outcomes. Despite these achievements, challenges remain concerning computational efficiency, training complexity, and generalization to irregular and extensive missing regions. Future research directions identified include improving model efficiency for ultra-high-resolution tasks, strengthening global semantic coherence by incorporating vision-language priors, enhancing user controllability via multi-modal inputs, and developing better perceptual evaluation metrics. This paper systematically reviews state-of-the-art Transformer-based and Diffusion-based methods, analyzes their strengths and limitations, and outlines critical areas for further advancement, providing valuable insights for ongoing research in image inpainting.

Keywords: Image inpainting; transfomer; diffusion Model.

1. Introduction

With the rapid advancement of artificial intelligence, deep learning has become a significant driving force in the field of image processing. Image inpainting, a key task in computer vision, aims to restore damaged or obscured areas of an image, ensuring visual coherence and authenticity. Recently, large deep learning models, especially Transformers and Diffusion Models, have demonstrated remarkable performance in image inpainting tasks, attracting widespread attention from both academia and industry.

Traditional image inpainting methods generally rely on texture matching and diffusion strategies based on partial differential equations [1, 2]. These methods perform well on small damaged areas with simple textures but often struggle to maintain realism and visual quality in complex scenes or large damaged regions. The introduction of deep learning methods, such as Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), marked significant progress by improving texture quality and local detail recovery [3, 4]. However, CNNs typically lack sufficient global context understanding, and GAN-based approaches often encounter issues related to training instability and structural coherence, limiting their effectiveness in complex or large-region inpainting tasks.

The emergence of Transformer-based models and the rapid development of Diffusion Models have led to further breakthroughs in the image inpainting domain. Transformer models leverage powerful global feature extraction capabilities and attention mechanisms, effectively capturing long-range dependencies and significantly enhancing visual consistency in repaired regions. Models such as Vision Transformer (ViT) and Swin Transformer have made notable advancements, particularly in handling large and irregularly shaped missing areas [5, 6]. Additionally, Diffusion Models

progressively restore damaged areas through iterative denoising steps, significantly enhancing realism and the preservation of texture details. This fine-grained restoration approach shows substantial advantages, particularly for high-resolution images and complex texture scenarios.

Despite their impressive performance, Transformer-based and Diffusion-based methods still face several challenges, including high computational requirements, training inefficiency, and difficulties generalizing effectively to irregular or extensive damaged regions. Therefore, further research and optimization of Transformers and Diffusion Models in image inpainting have significant theoretical and practical importance.

This paper focuses primarily on Transformer-based and Diffusion Model-based methods due to their superior capability in modeling global contexts and generating diverse, high-quality results, especially suitable for complex inpainting scenarios. The author provides a detailed discussion of these methods, analyzing their strengths and limitations, and exploring future development directions, thus offering a comprehensive reference for further research in this field.

2. Transformer-Based Image Inpainting Methods

2.1. Exploration of Transfomer-based Inpainting

Transformers achieved initial success in natural language processing and have been introduced to computer vision in recent years. The Vision Transformer (ViT), for example, divides an image into patches and processes them like a sequence of words, using self-attention to encode information from the entire image [7]. Unlike convolutional networks, the self-attention mechanism in Transformers can directly model long-range dependencies between arbitrary pixels. This capability is highly suited to the needs of image inpainting: for a large missing region, the model must utilize clues from distant parts of the image to infer reasonable content. Traditional CNNs, limited by their receptive field, struggle to effectively use far-away context; in contrast, a Transformer can "see" every part of the image globally through self-attention, giving it a natural advantage in understanding global structure and maintaining semantic consistency. Transformers thus hold promise for addressing scenarios that CNNs and GANs find difficult – such as ensuring that the filled content is semantically consistent with the rest of the image and does not violate the overall structure.

2.2. Pure-Transformer Architectures in Inpainting

Some studies have explored using purely Transformer-based architectures for image inpainting. For example, the Image GPT model by Mark Chen takes a sequence of image pixels as input and uses an autoregressive Transformer to generate the image pixel-by-pixel [8]. Essentially, such a model can predict missing pixels step by step given the context pixels, thereby performing image completion. However, since a pure Transformer incurs extremely high computational cost for high-resolution image generation, early explorations typically operated at lower resolutions or incorporated additional tricks. For instance, Niki Parmar trained a self-attention model on small images for image completion, demonstrating that a Transformer can capture image structures, albeit under limited resolution [9]. In general, while a fully Transformer-based model can maximally model global dependencies, applying it directly to high-resolution image inpainting still faces very large computational costs and data requirements.

2.3. Hybrid Models Combining Transformer and Convolution

Given the limitations of pure-Transformer models, many recent works have explored combining Transformers with convolutional networks to leverage the strengths of both. A prime example is MAT (Mask-Aware Transformer) proposed by Wenbo Li, designed specifically for large-hole image inpainting [10]. MAT introduces a mask-aware Transformer block: during self-attention, it aggregates global information only among tokens corresponding to valid pixels (non-missing regions), thus reducing interference from and computation on invalid (masked) regions. Meanwhile, MAT

fuses Transformers with a U-Net style convolutional architecture: overall it uses a U-Net-like encoder–decoder structure, but employs multi-layer Transformer encoders for high-level feature extraction, while still using convolutional decoding and upsampling for detail reconstruction. This hybrid design uses Transformers to ensure global semantic consistency, then relies on the convolutional network to restore local details and improve efficiency at high resolutions. MAT achieved state-of-the-art results at the time on large-mask inpainting tasks for datasets like Places2 and CelebA-HQ, demonstrating the effectiveness of combining Transformers with convolutions. Moreover, the idea of integrating Transformers with U-Net has seen success in other vision tasks and has been applied to image inpainting. For example, UNETR and Swin-Unet integrate Vision Transformers into a U-Net architecture for medical image segmentation, showing that combining the global features extracted by Transformers with the fine localization provided by U-Net can improve performance [11, 12]. Similarly, in low-level vision tasks, analogous explorations exist: for instance, SUNet replaces convolutional modules in U-Net with Swin Transformer blocks for image denoising, enhancing global modeling capability while retaining U-Net's efficiency [13]. These hybrid strategies indicate that combining the global attention of Transformers with the local detail handling of convolutions is beneficial for tasks like image inpainting and denoising.

2.4. Structure-Guided Transformer Models

The most recent Transformer-based inpainting methods have made improvements in leveraging structural guidance and enhancing model stability. Cao proposed ZITS++, an improved version of the earlier structure-guided inpainting model ZITS [14]. The original ZITS used a progressive Transformer to recover the overall image structure in a low-resolution sketch space, then coupled it with a CNN-based texture generator to separately reconstruct structure and texture. ZITS++ further enhances the Transformer's stability and capability for inpainting: it introduces techniques like zero-initialized residual addition (ZeroRA) to efficiently fuse structural priors, and improves the mask positional encoding strategy, enabling more reasonable structural reconstructions when dealing with large missing regions. Compared to its predecessor, ZITS++, on datasets like Places2, significantly improves structural completeness and detail realism in complex scenes, yielding more stable and reliable filling in regions with high structural uncertainty.

Wu proposed SyFormer (Structure-Guided Synergism Transformer), designed for images with a large missing ratio [15]. To address the inconsistent feature representations and lack of structural cues caused by extensive missing regions, SyFormer introduces a dual-routing filtering module that employs a progressive filtering strategy to eliminate invalid noise and interference, while simultaneously establishing global texture correspondences during the inpainting process. This structure-guided Transformer synergistically fuses global semantic signals with local detail cues, effectively mitigating semantic discrepancies and enriching structural information for high-resolution inpainting with large holes. Experimental results show that in heavily occluded scenarios, SyFormer produces images with much more reasonable structure and convincing realism, with notable improvements in metrics compared to previous Transformer-based models.

2.5. Advantages and Limitations of Transformer-Based Inpainting

Transformer-based methods offer significant advantages for image inpainting, primarily through their robust capability for global context understanding. Leveraging self-attention mechanisms, Transformers effectively capture long-range dependencies within an image, facilitating semantically coherent and consistent inpainting results even for large missing regions. For instance, in landscape photos with extensive gaps, Transformers can utilize distant contextual clues to seamlessly reconnect fragmented structures or maintain coherent scene composition. Structure-guided Transformer approaches, such as ZITS++ and SyFormer, further reinforce global structural consistency, enhancing the overall effectiveness of these models.

Hybrid Transformer-CNN architectures exploit complementary strengths of both global attention and local convolution, ensuring that inpainted regions align with global scene semantics while

maintaining sharp local textures. This integration typically achieves higher fidelity compared to models relying solely on CNNs or Transformers.

However, Transformer-based models present certain challenges, notably their computational complexity and intensive memory requirements. The self-attention mechanism scales quadratically with token count, posing significant constraints for high-resolution image processing. Despite optimizations such as sparse or selective attention mechanisms (e.g., MAT), Transformer models remain resource-intensive and necessitate extensive training datasets to achieve robust generalization. Consequently, pure Transformer architectures can be less effective when trained on limited datasets, often requiring supplementary CNN decoders or additional loss functions (e.g., adversarial or perceptual losses) to enhance fine texture synthesis, a recognized limitation of standalone Transformer models. Approaches like MAT and ZITS++ explicitly address this limitation by incorporating modules dedicated to refining high-frequency details.

In summary, Transformer-based image inpainting methods are particularly advantageous for scenarios involving extensive missing areas that demand comprehensive image understanding. Their use may be suboptimal for simpler, small-gap scenarios efficiently handled by lightweight CNNs. Continued research is directed toward enhancing Transformer efficiency and refining hybrid architectures, aiming for improved computational performance without compromising their powerful global reasoning capabilities.

3. Diffusion Model-Based Image Inpainting Methods

3.1. Concept and Rise of Diffusion Models

Diffusion models are a class of generative models based on progressively perturbing data with noise. Their core idea is inspired by non-equilibrium thermodynamics: gradually add noise to the data until it becomes pure noise, while simultaneously training a model to learn the reverse process, i.e. gradually denoise to recover the data [16]. The DDPM (Denoising Diffusion Probabilistic Model) proposed by Ho marked a breakthrough for diffusion models in image generation [17]. They showed that with a sufficient number of diffusion steps, diffusion models can generate images of quality comparable to GANs while covering the data distribution more completely. Subsequently, Song improved the sampling process of diffusion models by proposing DDIM (Denoising Diffusion Implicit Models), which drastically reduced the number of sampling steps and significantly boosted generation efficiency [18]. In the past three years, diffusion models have risen rapidly across various image generation domains, not only surpassing GANs in unconditional image generation but also being successfully applied to tasks like image translation and image editing [19].

3.2. Conditional Diffusion Models for Inpainting

The strong generative capability of diffusion models has drawn researchers to apply them to image inpainting. The iterative generation mechanism of the diffusion process can be leveraged to repeatedly sample and denoise the missing region until content that blends with the known regions is produced. The key to achieving this is incorporating conditional constraints into the diffusion model to ensure that known pixels of the image remain unchanged while only the unknown region is synthesized. For example, RePaint proposed by Lugmayr was one of the pioneering works applying diffusion models to image inpainting [20]. Built on a DDPM foundation, their method performs repeated diffusion sampling for the missing region during inference: it initializes the unknown region with random noise while keeping known regions fixed, then alternates between the reverse diffusion denoising steps of the model and re-injecting the known pixels. After multiple iterations, the missing region gradually converges to content consistent with the context. This method exploits the strong generative power of diffusion models and can produce diverse, high-fidelity inpainting results even for the same input mask.

In addition to RePaint, the Palette model by Saharia uses a conditional diffusion framework to achieve high-quality results on multiple image-to-image translation tasks with a single model, including inpainting, colorization, and deblurring [21]. OpenAI's GLIDE (Nichol et al., 2021) and the subsequent DALL·E 2 further conditioned diffusion models on text, guiding the inpainting of missing regions according to textual descriptions. Another important development was the latent diffusion model proposed by Rombach: they perform the diffusion process in a low-dimensional latent space of the image (instead of in pixel space), drastically reducing the computational cost of high-resolution image synthesis while using a pretrained encoder to preserve the detail quality of generated images [22]. Latent diffusion was employed in the well-known Stable Diffusion model, providing powerful image editing and inpainting capabilities. The introduction of diffusion models has brought unprecedented performance improvements to image inpainting, greatly enhancing the diversity and realism of the generated results.

3.3. Structure-Guided Diffusion Models

The latest research on diffusion models focuses on enhancing user control and structural constraints. Xie proposed SmartBrush, a multi-modal diffusion model that enables user-guided image inpainting and object insertion [23]. Unlike general diffusion models (e.g., Stable Diffusion) which rely solely on text prompts, SmartBrush conditions on both a text prompt (describing the semantics of the desired object) and a shape mask (specifying the outline of the object to generate). The method adds an object mask prediction branch to the diffusion U-Net, and through specialized training and sampling strategies, it strengthens the protection of background regions to prevent the generated object from altering surrounding textures. Moreover, SmartBrush employs multi-task joint training on image inpainting and text-to-image generation, leveraging large-scale text–image data to boost the model's generative capability. Experiments on COCO demonstrate that SmartBrush outperforms existing baselines in terms of image quality, shape control accuracy, and background preservation.

Liu proposed StrDiffusion (Structure-Guided Diffusion Model) to address semantic inconsistencies in diffusion-based inpainting by leveraging structural information [24]. In conventional diffusion inpainting, noise is added only to texture in the forward process and the missing region is eventually reduced to pure noise, which can create a "semantic gap" between preserved and missing regions. To overcome this, StrDiffusion incorporates structural constraints into the diffusion process: it trains a structure prediction network to provide sparse structural guidance, ensuring that structural features inside and outside the mask remain aligned even in the early denoising stages. The model adaptively determines when to shift from structure-guided generation to texture generation, dynamically balancing their contributions. Experiments show that on benchmarks like Places2, StrDiffusion achieves inpainting results with better global semantic coherence and detail realism than existing diffusion models, effectively narrowing the semantic gap between the filled region and the background.

3.4. Advantages and Limitations of Diffusion-Based Inpainting

Diffusion-based inpainting methods have achieved state-of-the-art performance primarily due to their ability to generate high-quality, realistic, and diverse outputs. Through iterative denoising, diffusion models effectively capture intricate details and textures, surpassing many GAN-based approaches in perceptual quality. Their stochastic nature allows for multiple plausible completions from the same input, beneficial in creative and restoration scenarios.

Furthermore, diffusion models exhibit stable training dynamics, avoiding common GAN issues such as mode collapse, and can readily integrate various conditioning inputs like masks, text, or sketches, enabling precise and flexible inpainting control.

However, diffusion methods face limitations, notably high computational cost due to iterative inference processes, making them slower than CNN or GAN models. For instance, generating a 512×512 image with 50 diffusion steps means 50 passes through a U-Net, which can take several

seconds on a GPU – this is fine for offline tasks but too slow for real-time applications. Although algorithms like DDIM and more recent advancements have reduced the required steps (sometimes one can get away with as few as 20-30 steps), the speed gap remains significant [25].They may also produce semantically inconsistent results if not adequately guided, requiring additional structural guidance or context-aware strategies. Weak conditioning can lead to boundary inaccuracies, necessitating methods like pixel resetting for correction.

In summary, diffusion-based inpainting excels in applications prioritizing realism and diversity but remains challenging for real-time or resource-limited scenarios. Current research continues to improve computational efficiency and coherence through model optimizations and hybrid approaches.

4. Common Datasets in Image Inpainting Research

General Scene Image Datasets: Common general-purpose scene datasets include Places2 and COCO [26, 27]. Places2, released by MIT, is a large-scale dataset covering various categories such as natural landscapes, city streets, and indoor scenes. In image inpainting (especially image completion), Places2 is often used as a training set due to its rich contextual diversity, helping train deep models to fill large holes effectively. Recent Transformer- and diffusion-based methods, such as TransFill, trained on Places2 have significantly outperformed previous approaches. Places2 provides data suitable for learning both global semantics and fine texture details. The Places365-Standard version contains approximately 1.8 million training images across 365 scene categories, while the extended version includes about 10 million images across 434 categories [28]. Most images exceed 256px in resolution without additional pixel-level annotations, and missing regions are typically simulated by applying random masks. The COCO dataset is also used for object-removal inpainting tasks due to its wide diversity of objects and backgrounds, though it requires additional preprocessing, such as random mask generation, primarily serving as a measure of model generalization.

Face Image Datasets: Common face datasets include CelebA/CelebA-HQ and FFHQ (Flickr-Faces-HQ) [29, 30, and 31]. CelebA, released by The Chinese University of Hong Kong, contains 202,599 celebrity face images with 40 attribute labels per image. CelebA-HQ is a high-quality subset curated by NVIDIA from CelebA, containing 30,000 face images with resolutions up to 1024×1024. These datasets are widely used for face inpainting tasks such as filling missing facial regions and face super-resolution. Due to their high resolution and quality, CelebA-HQ frequently serves as a face prior in diffusion-based models to generate realistic facial details. FFHQ, another high-quality face dataset from NVIDIA, includes 70,000 diverse faces at 1024×1024 resolution, covering various ages, ethnicities, backgrounds, and accessories, commonly utilized as a face prior in diffusion-based face generation.

Object Image Datasets: Specific object-category datasets used in inpainting research include the Facade dataset, containing 606 rectified building facade images from various cities and architectural styles; the Describable Textures Dataset (DTD), consisting of 5,640 texture images from nature, categorized into 47 classes; and the Stanford Cars dataset, providing 16,185 car images across various makes, models, colors, and viewpoints, along with detailed model annotations [32, 33]. These datasets, although not originally designed for inpainting, are useful for evaluating a model's capability to handle missing content of specific object types.

Video Inpainting and Enhancement Datasets: In video domains, datasets used for video completion include YouTube-VOS, DAVIS, Vimeo-90K, and REDS [34, 35, 36, 37]. YouTube-VOS, introduced for video object segmentation, contains 3,471 training videos and 508 test videos with diverse scenes and accompanying foreground object masks, suitable for testing video inpainting tasks involving occlusions and significant motion. DAVIS 2017 includes 150 high-quality videos with pixel-level segmentation annotations, commonly utilized for video completion evaluation. Vimeo-90K provides rich data for low-level video processing tasks: its subsets include "tri-frame sequences" (73,171 sequences for interpolation) and "seven-frame sequences" (91,701 sequences for denoising, deblocking, and super-resolution). The REDS dataset, presented in NTIRE 2019, comprises 300

video sequences (100 frames each, at 720p resolution) for challenges in deblurring and super-resolution. These datasets assess the performance and temporal consistency of image inpainting techniques extended to video data.

5. Conclusion

Despite progress in image inpainting, critical challenges persist. Future research should prioritize:
Efficient Ultra-High-Resolution Processing: Reduce computational costs of Transformers and diffusion models through lightweight architectures (e.g., compressed blocks), accelerated sampling (fewer steps or parallel denoising), and model distillation for real-time performance.

Semantic Coherence in Complex Scenes: Integrate vision-language models (e.g., CLIP) or knowledge graphs to enhance global scene understanding, ensuring inpainted content aligns with contextual semantics and avoids logical inconsistencies.

Multi-Modal User Interaction: Enable flexible control via sketches, text prompts, or reference images while preserving seamless visual harmony with the original image.

Perceptual Quality Optimization: Develop metrics that better quantify semantic plausibility and visual realism (e.g., learned perceptual scores) and design training objectives aligned with human subjective evaluations.

In summary, future research in image inpainting is likely to pursue methods that are smarter, more efficient, and more controllable. On one hand, it will be important to integrate diverse sources of information and priors (semantics, structures, user intent, etc.) to improve inpainting quality and plausibility. On the other hand, balancing model complexity with practical deployment needs will be crucial to overcome computational bottlenecks for large-scale applications. With deeper integration of generative models and semantic understanding, we can expect breakthroughs in performing inpainting in more complex real-world scenarios, achieving results that are even more realistic and imperceptible to human observers.

References

[1] Bertalmio M, Sapiro G, Caselles V, Ballester C. Image inpainting. Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH '00). ACM Press/Addison-Wesley Publishing Co., USA, 2000:417–424.

[2] Criminisi A, Perez P, Toyama K. Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on Image Processing. 2004, 13(9):1200–1212.

[3] Zhang Ye & Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820. 2015.

[4] Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Advances in neural information processing systems. 2014, 27.

[5] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.

[6] Liu Ze, Lin Yutong, Cao Yue, et al. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF international conference on computer vision. 2021, 10012-10022.

[7] Dosovitskiy A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020.

[8] Chen M, Radford A, Child R, et al. Generative pretraining from pixels. Proceedings of the 37th International Conference on Machine Learning (ICML'20). 2020, 119:1691-1703.

[9] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer. International conference on machine learning. 2018, 4055-4064.

[10] Li Wenbo, et al. Mat: Mask-aware transformer for large hole image inpainting. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[11] Hatamizadeh A, et al. Unetr: Transformers for 3d medical image segmentation. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.

[12] Cao Hu, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. European conference on computer vision. 2022.

[13] Fan C-M, Liu T-J, Liu K-H. SUNet: Swin Transformer UNet for Image Denoising. 2022 IEEE International Symposium on Circuits and Systems (ISCAS). 2022, 2333-2337.

[14] Cao Chenjie, Dong Qiaole, Fu Yuanwei. ZITS++: Image Inpainting by Improving the Incremental Transformer on Structural Priors. IEEE Trans Pattern Anal Mach Intell. 2023, 45(10):12667-12684.

[15] Wu Jie, Feng Yuchao, Xu Honghui, et al. SyFormer: Structure-Guided Synergism Transformer for Large-Portion Image Inpainting. Proceedings of the AAAI Conference on Artificial Intelligence. 2024, 38(6):6021-6029.

[16] Sohl-Dickstein J, et al. Deep unsupervised learning using nonequilibrium thermodynamics. International conference on machine learning. 2015.

[17] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems. 2020, 33:6840-6851.

[18] Song Jiaming, Meng Chenlin, Ermon S. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502. 2020.

[19] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. Advances in neural information processing systems. 2021, 34:8780-8794.

[20] Lugmayr A, et al. Repaint: Inpainting using denoising diffusion probabilistic models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[21] Saharia C, et al. Palette: Image-to-image diffusion models. ACM SIGGRAPH 2022 conference proceedings. 2022.

[22] Rombach R, et al. High-resolution image synthesis with latent diffusion models. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[23] Xie S, et al. Smartbrush: Text and shape guided object inpainting with diffusion model. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.

[24] Liu Haipeng, Wang Yang, Qian Biao, et al. Structure Matters: Tackling the Semantic Discrepancy in Diffusion Models for Image Inpainting. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024, 8038-8047.

[25] Shih A, Belkhale S, Ermon S, et al. Parallel sampling of diffusion models. Advances in Neural Information Processing Systems. 2023, 36: 4263-4276.

[26] Zhou Bolei, et al. Places: An image database for deep scene understanding. arXiv preprint arXiv:1610.02055. 2016.

[27] Lin T-Y, et al. Microsoft coco: Common objects in context. Computer vision–ECCV 2014. 2014.

[28] Jing Longlong, Tian Yingli. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence. 2020, 43(11):4037-4058.

[29] Zhang Honglun, et al. Show, attend and translate: Unpaired multi-domain image-to-image translation with visual attention. arXiv preprint arXiv:1811.07483. 2018.

[30] Huang Huaibo, et al. Introvae: Introspective variational autoencoders for photographic image synthesis. Advances in neural information processing systems. 2018, 31.

[31] Karras T, Laine S, Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[32] Zhang Yuanhan, et al. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. Computer Vision–ECCV 2020. 2020.

[33] Cimpoi M, et al. Describing textures in the wild. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.

[34] Xu Ning, et al. Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327. 2018.

[35] Zhel tukhin A A. Phenomenological Lagrangians, gauge models and branes. Physics of Particles and Nuclei Letters. 2017, 14: 312-317.

[36] Xue Tianfei, et al. Video enhancement with task-oriented flow. International Journal of Computer Vision. 2019, 127:1106-1125.

[37] Deng Yuefan, et al. Optimal low-latency network topologies for cluster performance enhancement. The Journal of Supercomputing. 2020, 76(12): 9558-9584.