Inpainting Portrait-Map-Art with Diffusion-based framework

THI-NGOC-HANH LE, International University, Viet Nam and VNU-HCM, Viet Nam
DONG-YI WU[†], National Cheng-Kung University, Taiwan, Republic of China
SHENG-YI YAO[†], National Cheng-Kung University, Taiwan, Republic of China
TONG-YEE LEE[*], National Cheng-Kung University, Taiwan, Republic of China

Image inpainting, an essential task within the broader field of image restoration, aims to reconstruct missing or corrupted regions in images, ensuring they blend seamlessly with their surroundings. We propose Diff-PMArt to focus on a specific and complex type of inpainting involving portrait-map-art (PMA) images, where artistic portraits are integrated within cartographic backgrounds. Our goal is to restore the map background after removing the portrait, ensuring that the reconstructed region harmonizes with the surrounding structure and visual patterns on the map. Our results demonstrate the ability of Diff-PMArt to produce high-quality, realistic inpainting for PMA images, significantly improving over previous methods in handling unknown regions. In contrast to existing solutions, our approach can inpaint unknown regions without annotations, addressing a more intractable problem where no prior knowledge of the mask or labels is available. The proposed framework is carefully designed with two modules: a Visual Object Extraction (VOE) module, which identifies and extracts the portrait region as a mask, and a Diffusion-based Image Reconstruction (DIR) module, which uses a diffusion process to progressively inpaint the masked region. Our experimental results and evaluations demonstrate that Diff-PMArt has ability to reconstruct visually realistic and semantically coherent background of PMA images.

CCS Concepts: • **Computing methodologies → Image manipulation**.

Additional Key Words and Phrases: PMA, map art, Diff-PMArt, diffusion, inpainting, multimedia

ACM Reference Format:
Thi-Ngoc-Hanh Le, Dong-Yi Wu, Sheng-Yi Yao, and Tong-Yee Lee. 2026. Inpainting Portrait-Map-Art with Diffusion-based framework. *J. ACM* , (2026), 24 pages. https://doi.org/10.1145/xxxxxxxxxxxxx

1 INTRODUCTION

Image restoration has recently become a pivotal area of research and application, encompassing various tasks aimed at enhancing and recovering visual content. These tasks include image super-resolution [17, 30, 50], deblurring [26, 64, 65], dehazing [23, 63, 71], and inpainting [13, 22, 27, 31, 32, 35, 40, 48, 56, 58, 60, 61]. Among these, image inpainting holds particular significance in the domain of multimedia communication. By enabling the seamless reconstruction of missing or corrupted regions in images, inpainting facilitates the creation of visually coherent and engaging content, which is critical for maintaining communication quality and aesthetic appeal. The rising demand

[*]Corresponding author.
[†]Dong-Yi Wu and Sheng-Yi Yao contributed equally to this work.

Authors' addresses: Thi-Ngoc-Hanh Le, International University, Ho Chi Minh City, Viet Nam and VNU-HCM, Viet Nam, ltnhanh@hcmiu.edu.vn; Dong-Yi Wu, National Cheng-Kung University, Tainan, Taiwan, Republic of China, cutechubbit@gmail.com; Sheng-Yi Yao, National Cheng-Kung University, Tainan, Taiwan, Republic of China, nd8081018@gs.ncku.edu.tw; Tong-Yee Lee, National Cheng-Kung University, Tainan, Taiwan, Republic of China, tonylee@mail.ncku.edu.tw.

0004-5411/2026/-ART $15.00
https://doi.org/10.1145/xxxxxxxxxxxxx

Fig. 1. Each pair: PMA by arttist (left), inpainted background of PMA by our method (right).

for sophisticated image manipulation, driven by the prevalence of social media and multimedia platforms, underscores the essential role of inpainting in modern communication ecosystems.

Image inpainting, which has the goal to reconstructing missing or corrupted areas in digital images, has seen significant advancements in recent years, particularly with deep learning techniques [8, 24, 32, 35, 40, 58, 60]. However, while traditional image inpainting techniques have made great strides in recovering missing or corrupted areas of natural images, their limitations become more evident when dealing with more complex, structured domains, such as streets, bridge, etc. One such domain is "*Portrait-Map-Art*" (PMA) images - a composite artistic form in which geographic map structures (*e.g.*, roads, rivers, and boundaries) are arranged to form a silhouette of human portrait, as shown in Fig.1 (left image in each pair). This art style has gained recognition through British portrait artist Ed Fairburn*. Some early research explores this modern art form by investigating models for transferring styles, with the aim of adapting to a wide variety of artistic styles [19, 41, 68]. Unlike natural images, PMA exhibits dense line patterns, high-frequency textures, and strict geometric alignment inherited from cartographic layouts. These characteristics make inpainting particularly challenging, as reconstruction must preserve both structural coherence and color harmony across portrait and map components.

Prior work in image inpainting has predominantly centered on restoring natural scenes, where texture synthesis and color blending are key considerations. However, PMA images present a fundamentally different set of challenges. The map backgrounds often consist of intricate geographical features - sharp boundaries, continuous contours, and gradient-like color transitions - which must be reconstructed in a coherent and visually seamless manner. Generally, the existing solutions can now produce compelling results on inpainting natural images. Nevertheless, they share two mutual points: (1) models learn from richly available paired data, (2) natural images are the data they work on. Consequently, adjusting the high-level semantics of image content while maintaining realistic visuals presents a greater challenge in current research efforts.

In contrast, our proposed framework represents a significant advancement by addressing a more challenging inpainting scenario: restoration does not require prior knowledge about the corrupted regions or masks, *i.e.,* blind inpainting [37]. Unlike prior approaches that rely on rectangular masks and labeled maps, our system is designed to handle more intractable and difficult scenarios where the mask is automatically extracted, and no labeled plain maps or annotated data are available. These requirements remarkably introduce a new level of difficulty, as distinguishing between trustworthy and unreliable regions for reconstruction becomes far more complex and challenging. The absence of such structured masks makes the task significantly harder, demanding a model that

*https://edfairburn.com/full-portfolio

can infer and recreate missing regions based on the intrinsic patterns of the map without explicit supervision.

To address these challenges, we propose Diff-PMArt, a framework specifically designed to tackle the complexities of PMA image inpainting. Diff-PMArt integrates two core modules: the Visual Object Extraction (VOE) module, which autonomously identifies and extracts the portrait region, and the Diffusion-based Image Reconstruction (DIR) module, which inpaints the portrait region using a diffusion strategy. The VOE module produces a mask that sufficiently covers the portrait, while the DIR module receives this mask and the original image, using diffusion processes to reconstruct the masked areas. By progressively refining the image through denoising, the diffusion model enables precise pixel generation that aligns with the geometric structure and color distribution of the surrounding mask areas. This process is trained on triplets comprising the input image, random masks, and edge maps, allowing the model to generalize well to unknown regions. At the core of Diff-PMArt is the use of diffusion-based techniques, a recent class of generative models for image synthesis and manipulation. Recently, diffusion models have gained recognition as a powerful generative approach in computer vision, particularly for tasks requiring the synthesis of high-quality, coherent content [15, 21, 31, 49, 53]. In Diff-PMArt, the DIR module utilizes these techniques to reconstruct the masked region of a PMA image by learning from a triplet of inputs: the original image, a randomly generated mask, and the edge map of the input. This enables our model to capture and replicate map-specific attributes such as color gradients and boundary coherence, ensuring that the reconstructed region blends seamlessly with the surrounding content.

In this paper, our experimental results and evaluations demonstrate that our system has ability to restore map backgrounds with high fidelity to the original structure, surpassing prior supervised methods in handling unannotated, and unknown regions. With this ability, our approach contributes valuable insights for broader applications in image editing. In summary, the key contributions of our work are as follows:

- We introduce the Diff-PMArt framework for PMA image inpainting application.
- We handle inpainting in an unsupervised manner, where unknown regions are reconstructed without annotations. This presents a more intractable and difficult challenge, particularly in distinguishing trustworthy from unreliable regions during reconstruction.
- Our framework offers broader implications, addressing the growing need for sophisticated image editing tools across fields such as digital art, social media, heritage restoration, and interactive visual content creation.

2 RELATED WORK

In this section, we discuss the overview of the techniques used in prior works which are related to our current research: image restoration and image inpainting.

Image Restoration. Image Restoration (IR) has long been a core focus in low-level vision research, essential for enhancing the perceptual quality of images. Common IR tasks include [17, 30, 50], deblurring [26, 64, 65], inpainting [22, 31, 32, 35, 40, 48, 56, 58, 60, 61], dehazing [23, 63, 71] and removing compression artifacts. Traditional IR methods approach restoration from a signal-processing perspective, often using hand-crafted algorithms in spatial or frequency domains to minimize artifacts. With advances in deep learning, extensive datasets tailored to specific IR tasks (such as deraining and motion deblurring) have been developed. Leveraging these datasets, recent research has emphasized enhancing IR network architectures, typically based on CNNs [18] or Transformers [46], to handle complex degradations. Although these methods have significantly improved objective image quality, challenges remain with realistic texture generation, limiting the applicability of IR in real-world settings [24].

Recently, diffusion models have enabled significant advancements in visual generation tasks [11, 43, 44]. Typically, a diffusion model comprises a forward (or diffusion) process and a reverse process: the forward process gradually adds pixel-level noise to an image until it resembles Gaussian noise. Inspired by the powerful generative capabilities of diffusion models, numerous studies have explored their application in image restoration tasks, with a focus on enhancing texture recovery [21, 33, 39, 54]. Saharia et al. [39] adapt denoising diffusion probabilistic models to conditional image generation and performs super-resolution through a stochastic iterative denoising process. Li et al. [21] reconstruct high-resolution (HR) images from the given low-resolution (LR). To speed up convergence and stabilize training, they introduce residual prediction by taking the difference between the HR and LR image as the input in the first diffusion step, making the model focus on restoring high frequency details. Özdenizci and Legenstein [33] introduce a patch-based diffusive image restoration algorithm that enables processing of images of any size using denoising diffusion models. Whang et al. [54] adopt a different perspective and view deblurring as a conditional generative modeling task, where we seek to generate diverse samples from the posterior distribution. However, these diffusion models face two main limitations: (1) training from scratch requires a large amount of paired training data, and (2) obtaining paired image annotation is challenging. These issues make unannotated map art image restoration increasingly difficult with diffusion models. For more related work, readers can see the comprehensive survey of diffusion models for image restoration and enhancement [24].

Image Inpainting Image inpainting relates to reconstructing missing areas in an image while ensuring the result is both visually realistic and semantically coherent. A tough problem of image inpainting is to restore complex structures in the corrupted regions [28]. Researchers explore challenges in this topic with several approaches, such as patch-based [1, 10], Deep Generative Models [16, 29, 40, 67]. And most recently, researchers explore this problem by diffusion-based models [62]. With the real-world applications increasing, more and more works also take "*where to inpaint*" into consideration, and a series of blind image inpainting methods have been proposed.

In blind image inpainting, the mask that indicates the location of corrupted regions is unknown. Therefore, the blind settings require an extra consideration about "*where to inpaint*" before "*how to inpaint*". To address these issues, researchers investigate several techniques [2, 47, 51]. Cai et al. [2] introduced the first blind image inpainting method using an end-to-end CNN architecture that directly learns to identify and restore corrupted regions. Wang et al. [51] developed a consistency network that first detects the regions for inpainting using predicted masks, followed by an inpainting network to fill them. Wang et al. [47] proposed a technique that predicts masks by incorporating contextual coherence and an additional frequency modality, then performs landmark prediction and ultimately inpaints facial images. Along with this direction, recent methods have also explored structural priors and contextual propagation for large-region inpainting. For instance, ZITS++ [4] leverages transformer-based structure restoration and Fourier-based texture reconstruction to recover global image structures under irregular masks, while E2FGVI [25] employs flow-guided feature propagation and content hallucination to maintain spatial consistency during large-area reconstruction. Recent works explore adaptive information selection to improve efficiency and representation quality, such as dynamic pruning in large language models [57] and unified static-dynamic temporal filtering for video grounding [14]. Although targeting different tasks, these approaches share the principle of selectively preserving informative signals while suppressing redundant content. Similarly, our Diff-PMArt performs spatially adaptive information selection via VOE-based masking and buffer construction, which regulates the structural cues provided to the diffusion model.

The sharp contrast between our framework and priors [2, 47, 51, 62] is its ability to inpaint artworks without relying on data annotation via diffuse strategy. Our method offers a comprehensive

solution, as it not only automatically detects *where to inpaint* but also addresses the *how* with precision, effectively tackling the challenges posed by the intricate and dense structures found in artworks.

3 METHODOLOGY

3.1 System Overview

We propose a diffusion-based framework for reconstructing the background of PMA images. Let $\mathcal{P}^o \in \mathbb{R}^{H\times W\times 3}$ be an input PMA image. The edited region is an irregular shape and is represented by a binary mask $m \in 0, 1^{H\times W}$ where a value of 1 indicates the editable positions in $\mathcal{P}^o$. We aim to generate an image from $\{\mathcal{P}^o, m\}$, such that the region where $m = 0$ remains as in $\mathcal{P}^o$, while the region where $m = 1$ aligns well with its surrounding pattern/structure and fits harmoniously. Our goal in this work is to automatically (1) predict m and (2) recover the structure of PMA's background such that the reconstructed image looks plausible and photo-realistic, see Fig.1 (right).

This task is both highly challenging and intricate, as it inherently involves a series of complex steps, such as determining *where to paint* and deciding *how to paint*. Firstly, the challenge falls in defining the editable positions as they are discrete regions rather than a rectangle one in prior work of natural image inpainting. Secondly, we are to inpaint unknown regions without knowledge about the annotations, which is more intractable and difficult than the supervised settings. The lack of masks and annotated plain maps makes distinguishing between trustworthy and unreliable regions difficult to reconstruct or recover. And thirdly, it's demanding to control the consistency not only for color but also edge attributes (*e.g.* line, road) between the inner and outer regions after reconstruction. To resolve these challenges, we design our framework (dubbed Diff-PMArt) with two modules, a visual objected extraction (VOE) and a diffusion-based image reconstruction (DIR). The specific procedure is outlined in Fig. 2.

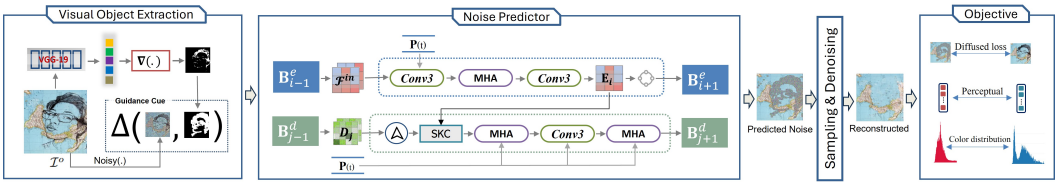


Fig. 2. Visualization of our system overview.

The VOE module shoulders the task of allocating the object region and extracting it as a mask. The extracted mask is acquired to be sufficient to cover the visual object fully. We model this module by defining $f_{VOE} : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W}$, with H and W are the height and width of the input image, respectively. We approach this function with a *fine-to-coarse* strategy. This technique can define the exact visual object region and the buffer area, which helps reconstruct blind patterns. The DIR module receives the extracted mask from f_{VOE} and the source PMA image as input, then formulates them with time embedding upon a diffusion technique. In this DIR module, we aim to produce a visually realistic background map image that mimics the color distribution in the input image and forms coherent edges. As shown in Fig.2, the DIR module uses a triplet of the input image, a random mask, and the edge map of the input image in the training manner. The training teaches the model how to inpaint the masked region with map attributes that align well with the outer areas. Once trained, given a PMA image, and the mask produced by our VOE module, our trained DIR model can efficiently and effectively reconstruct a clean and realistic background map image. Later, we will experimentally evaluate several challenging examples of PMA images.

3.2 Visual Object Extraction

The definition of visual object mask is essential for background reconstruction of PMA images. It indicates the exact location of visual object regions to be reconstructed. Unlike defining object in natural images, accurately extracting mask in PMA images is more challenging as the visual object is tangles with the map-based background like they belong together. To define the so-called mask in PMAs, some possible solutions can be considered, such as salient object detection (SOD) [20, 38, 70], bounding box detection (BBD) [6], or camouflaged object detection (COD) [7, 55]. However, object (which is also referred to as *"portrait"*) in PMA images is tangled with such dense and complex structure of the background. The off-the-shelf models of SOD, BBD, COD are not able to estimate precisely as they are not trained on PMA images. Alternately, we could devise a new network and train it on a self-prepared PMA dataset with a hope to generate proper masks. Nevertheless, this solution may face two obstacles. First, the available data of PMA images is not so many to be sufficient for training a neural network. Second, the mask extraction is just a pre-processing phase but requiring many resources, therefore, it may make the overall system cumbersome and downgrade the system's compactness with marginal effectiveness. Using bounding box to cover the portrait region is also a simple and possible solution. However, this naive technique is not appropriate to our current data caused by two challenges. First, inpainting region is discrete sub-regions, it will eventually damage free regions. Second, in the scenarios that the patterns belonging to the portrait distributed discretely or even entirely PMA image, we need to solve the problem of reconstructing the entire image. For example, Fig. 1 (the image pair on the left), since the information of the portrait distributed discretely in a relatively large region, using bounding box certainly leads to inpainting entire image as the pattern is distributed over entire image.

Motivated by the above observations, we attempt to investigate a technique to describe the mask in PMA images. With this design, we aim to automatically define the portrait region with less resource rather than giving this burden to users. Prior work in the realm of image reconstruction, which might be relatively close to our goal such as image inpainting/hole filling, often approaches this problem via "*coarse-to-fine*" strategy. That is, a bounding box (coarse), which covers the object, is detected in advance. Thereafter, the box is gradually refined to match the shape/contour of object. In contrast, we approach this problem by defining a *"fine-to-coarse"* mask. The generated mask, abbreviated as F2C-Mask, is "fine" enough to accurately allocate the visual objected region and "coarse" enough to completely cover both object and its neighboring pixels. To do this, we elaborate as follows.

Given a PMA image $\mathcal{P}^o$, we leverage pre-trained model to learn the abstraction of $\mathcal{P}^o$. This idea is inspired from Shih et al. [41] that extracts the portrait in PMA images for style transfer. However, their technique is not stable for all cases and heavily depends on performance of the setting of selected filter. In contrast, we aim to investigate an auto-defined strategy that could tolerate to the diversity of PMA images. Shih et al. [41] extract the portrait in PMA by applying a kernel to filter the input PMA then comparing it to 64 feature maps extracted from 5^{th} layer of VGG model. This strategy may have some drawbacks as the users need to adjust the filter to find the best fit parameters according to the input. Unlike their strategy, which applies filter to the input and compares to 64 feature maps to manually generate the extracted portrait, we formulate them in an optimization manner to produce one image.

The activation maximization [69] is then employed to visualize and understand the features that a particular neuron or a group of neurons in a neural network are responding to. The goal is to generate input patterns that maximize the activation of specific neurons, providing insights into what the network has learned. Mathematically, this can be expressed as:

$$\mathcal{M}^F = \mathcal{M}^F + \alpha \nabla_{\mathcal{P}^o}(\Phi_k) \tag{1}$$

Fig. 3. (a) input image,(b) fine mask $\mathcal{M}^F$, (c) F2C-mask $\mathcal{M}^c$, and (d) buffer area formed by $\mathcal{M}^c$.

where α is the learning rate, Φ_k is the selected layer of activation that we are to optimize, and $\nabla_{\mathcal{P}^o}$ represents the gradient with respect to the input. We found that in PMA images, the visual objected regions are formed with more intensive color by pencil drawing. Inspired by this observation, we maximize the activation of the feature maps at 5^{th} layer of the pre-trained model. we opt $\Phi_k = 5$ to maximize the activation. It's worth noting that opting other layers for equation (1) is possible. However, as studied in Gatys et al. [9], earlier layers may lack of desired features in our current application. Noting here that this strategy is not general enough for natural images such as camouflaged images. With equation (1), we can obtain a so-called "fine-mask" ($\mathcal{M}^F$), which only contains the portrait region and eliminates background information (*e.g.,* map attributes). See Fig. 3-(b). Thereafter, we base on the fine-mask to thread it to form a coarse-mask ($\mathcal{M}^c$). The definition of $\mathcal{M}^c$ is performed as:

$$\mathcal{M}^c = g(\mathcal{M}^F, \kappa), \tag{2}$$

here, g is the dilation function, and κ is the kernel for the dilating. In our experiments, we empirically adjust κ in range of 20 to 30. However, depending on the specific portrait region in PMA, users can adjust κ to obtain a sufficiently decent mask. The dilation strategy in this equation serves two benefits. The $\mathcal{M}^F$ may not capture the portrait perfectly, it will thread to x and y-axis to fine-tune the region. On the other hand, we use this operation to create a so-called *buffer area*, see the visualization in Fig.3(d), along the shape of portrait. We formally define the buffer area Θ as the difference between the coarse and fine mask with the set of pixels x as follows:

$$\Theta = \{x \mid x \in \mathcal{M}^c \wedge x \notin \mathcal{M}^F\}. \tag{3}$$

The buffer area Θ represents a narrow boundary band surrounding the portrait region. This region preserves edge transitions and local color statistics near the object boundary, providing contextual constraints during the diffusion reconstruction process. By explicitly modeling this transitional zone, the diffusion model receives additional structural guidance to enforce geometric continuity and suppress boundary artifacts. This effect is verified later in Section 4 via discussion and visualization of ablated results. Noting here that the proposed VOE module does require manually annotated portrait masks. Instead, it extracts visual structures via activation maximization on the pre-trained feature extractor, without additional mask-level supervision or task-specific fine-tuning. While this design enables annotation-free mask generation, its effectiveness depends on sufficient structural contrast between portrait elements and map textures, which may influence generalization to images with ambiguous boundaries or strong texture overlap.

3.3 Diffusion-based Image Reconstruction

With the extracted F2C-Mask, our goal in reconstruction phase is inpainting the masked region to recover the background visually realistic and geometrically consistent. To do this, we design a diffusion-based model. In a common diffusion-based model (DBM), they first progressively corrupt the input image x_0 to a noisy version $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I})$. The noise predictor mainly learns denoising ability, which is modeled as a Gaussian process with Markovian structure, *i.e.,* $\mu_\theta(x_t, t)$, while the reverse process aims to recover the data distribution from the Gaussian noise, $p_\theta(x_{t-1}|x_t)$. Alike this baseline workflow, we also model our DIR with this trait. However, directly applying forward and backward formulations as those in prior DBMs is trivial. We certainly face two obvious challenges: (1) clean image pairs are inaccessible and even unavailable in the real world, (2) PMA image data composes of rich attributes with various degradations. Consequently, to alleviate these challenges, we formulate our diffusion process as follows.

To address these challenges, we adopt the standard Denoising Diffusion Probabilistic Model (DDPM) formulation for the forward corruption process, while modifying the reverse denoising stage through a guidance-conditioned mechanism. Specifically, given a PMA image $\mathcal{P}^o$, the forward diffusion process follows the standard Markovian Gaussian corruption:

$$q(\mathcal{P}_t|\mathcal{P}_{t-1}) = \mathcal{N}\big(\sqrt{\alpha_t}\mathcal{P}_{t-1}, (1-\alpha_t)\mathbf{I}\big), \tag{4}$$

which yields the closed-form representation:

$$\mathcal{P}_t = \sqrt{\overline{\alpha}_t}\mathcal{P}^o + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}). \tag{5}$$

Importantly, the forward diffusion process remains identical to the standard DDPM formulation. Our contribution lies in conditioning the reverse denoising model on an additional guidance cue.

To provide explicit structural and contextual information for inpainting, we construct a guidance cue $\mathcal{S}$ defined as:

$$\mathcal{S} = (\mathcal{M}^c \odot \mathcal{P}^o) \oplus \tilde{\mathcal{P}}_t, \tag{6}$$

here, $\oplus$ denotes channel-wise concatenation, $\mathcal{M}^c$ is the complement of the inpainting mask, $\odot$ represents element-wise multiplication, and $\tilde{\mathcal{P}}_t$ is a noisy version of the original image sampled under the same diffusion schedule at timestep t. We call $\mathcal{S}$ in the term of *"guidance cue"*. Unlike conventional diffusion-based inpainting methods that condition solely on masked inputs, the proposed guidance cue jointly encodes (i) spatial mask structure and surrounding buffer regions and (ii) global noise statistics of the image. This provides the reverse denoising network with both structural priors and stochastic content, enabling more stable and geometrically consistent reconstruction.

Our goal in the reverse diffusion process is to progressively recover a clean reconstruction from the noisy sample $\mathcal{P}_t$. While the true reverse distribution $q(\mathcal{P}_{t-1} \mid \mathcal{P}_t)$ is intractable, DDPM approximates it using a parameterized Gaussian model. In our framework, we extend this formulation by conditioning the reverse process on the guidance cue $\mathcal{S}$. Specifically, we model the reverse transition as:

$$p_\theta(\mathcal{P}_{t-1} \mid \mathcal{P}_t, \mathcal{S}) = \mathcal{N}\big(\mu_\theta(\mathcal{P}_t, \mathcal{S}, t), \sigma_t^2\mathbf{I}\big), \tag{7}$$

where σ_t^2 follows the predefined variance schedule and $\mu_\theta(\cdot)$ is derived from a guidance-conditioned noise predictor.

Let $\epsilon_\theta(\mathcal{P}_t, \mathcal{S}, t)$ denote the neural network that estimates the added Gaussian noise at timestep t, following the standard DDPM parameterization, the mean of the reverse process is computed as:

$$\mu_\theta(\mathcal{P}_t, \mathcal{S}, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathcal{P}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathcal{P}_t, \mathcal{S}, t)\right). \tag{8}$$

Compared to the classical DDPM formulation where the noise predictor depends only on $(\mathcal{P}_t, t)$, our model conditions the estimation additionally on $\mathcal{S}$, thereby incorporating explicit structural and contextual priors into the denoising process.

The training objective of the model is to predict the noise using the objective:

$$\mathcal{L}_d = \mathbb{E}_{\mathcal{P}^o,\epsilon,t}\left[\| \epsilon - \epsilon_\theta(\mathcal{P}_t, \mathcal{S}, t) \|_2^2\right], \tag{9}$$

where $\mathcal{P}_t$ is generated from $\mathcal{P}^o$ under the forward diffusion schedule and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. In the coming subsection, we delve into detail how our noise predictor is modeled.

3.4 Noise predictor modeling

With the given the guidance cue $\mathcal{S}$ and noisy sample $\mathcal{P}_t$, our objective is to model the guidance conditioned noise estimator $\epsilon_\theta(\mathcal{P}_t, \mathcal{S}, t)$ introduced in Sec. 3.3. This network approximates the Gaussian noise added at timestep t and parameterizes the reverse transition distribution. Following standard diffusion architectures, we adopt an encoder-decoder U-Net backbone due to its strong multi-scale feature aggregation capability and effectiveness in noise prediction tasks. However, PMA images exhibit dense structures and complex degradations, which challenge the stability of a vanilla U-Net design. To address this, we introduce several architectural modifications tailored to our reconstruction scenario. Specifically, tweaks in our design are as follows.

As diagrammed in Fig.2, we design our noise predictor with five-layer depth in both encoder and decoder. To enable the model to attend to relevant features at different scales during both the contraction and expansion phases, we integrate Multi-Head Attention [5] (abbreviated as MHA) within both the encoding and decoding blocks, forming a *dual-Conv3* and *dual-MHA* protocols in encoding and decoding path, respectively. Each layer in the decoder is built by a so-called $\mathcal{B}^e$ block, which is structured by *Conv3* $\rightarrow$ MHA $\rightarrow$ *Conv3* $\rightarrow$ *MaxPooling*. The time embedding is incorporated to each $\mathcal{B}^e$ through the first *Conv3*. The presence of two convolutional layers (*Conv3*) before and after the MHA layer can enhance the model's ability to capture and refine fine details such as tiny roads in our map art application. The first *Conv3* can extract initial spatial features which are then contextualized by the attention mechanism, and refined by the second convolution before down-sampling. The second *Conv3* allows for an additional nonlinear transformation of the attention-modulated features, potentially leading to richer feature representations before the down-sampling step. This could be particularly beneficial in unsupervised settings where detailed feature extraction is essential for model performance. Formally, with the input feature maps $\mathcal{F}_i^{in}$ passed through a specific $\mathcal{B}^e$ is performed as:

$$\mathbf{E}_i = \mathcal{B}_i^e(\mathcal{F}^{in}) = Conv3\big(\xi(\mathcal{F}_{t.emb}^{in})\big), \tag{10}$$

with $\xi(.)$ represents for the MHA, $Conv3(.)$ indicates the standard convolution with kernel size of 3×3, and

$$\mathcal{F}_{t.emb}^{in} = Conv3(\mathcal{F}^{in}) + W_i T(t) \tag{11}$$

where W_i is weight matrix of the linear projection for i^{th} layer at time step t, designed to match the dimensionality of the features. Lastly, a *MaxPooling* operation is performed $MaxPooling(\mathbf{E}_i)$, which serves as input for the next block $\mathcal{B}_{i+1}^e$.

Let the tensor $\mathcal{F}_i^{out} \in \mathbb{R}^{h_i \times w_i \times k}$ denote the feature maps produced at each $\mathcal{B}_i^e$ with the above sequence; here k is the number of channels of each feature response, skip connection is then injected to facilitate information flow between different levels of the noise predictor and construct the decoded features. In the contracting path, we design decoder with blocks $\mathcal{B}^d$, which is structured by *UpSample* $\rightarrow$ SKC $\rightarrow$ MHA $\rightarrow$ *Conv3* $\rightarrow$ MHA. Given $\mathcal{D}_j$ as input feature maps at $\mathcal{B}_j^d$, skip

connection (SKC) is performed to connect $\mathcal{B}_i^e$ and $\mathcal{B}_j^d$ as:

$$\text{SKC}_j = UpSample(\mathcal{D}_j)\nabla\mathbf{E}_i, \tag{12}$$

here, $\mathbf{E}_i$ is the feature maps produced by Eq.(10) at block $\mathcal{B}_i^e$ such that $i + j = 5$, *i.e.,* the total layers in our design. The result of the skip connection is then fed to *dual-MHA*, which is mathematically expressed as:

$$\mathcal{D}_{j+1} = \xi\Big(Conv3\big(\xi(\text{SKC}_j) + \text{P}(t)\big) + \text{P}(t)\Big), \tag{13}$$

where $\text{P}(t) = W_{j+1}T(t)$ is the time embedding of time step t at layer $j + 1$. We note here that adding $\text{P}(t)$ after each significant operation (*i.e.,* MHA and *Conv3*) ensures the time-dependant context influences both local and global feature processing. This strategy furthermore enhances the model's capability to handle complex data structures and temporal dynamics in the diffusion process, making it particularly effective for our application which requires high fidelity in detailed reconstruction, such as in generating or denoising images with dense and complex attributes. The effectiveness of this design is verified via the ablated results in later section.

3.5 Network Optimization

Besides the loss function for diffusion process, *i.e.,* $\mathcal{L}_d$ in equation (9), we further optimize our model with other two objective functions, color consistency loss $\mathcal{L}_{cc}$ and content alignment loss $\mathcal{L}_{ca}$. It's worth noting that $\mathcal{L}_d$ alone is capable to generate results with decent visual effect. Let the network be denoted by f parameterized by the weights μ, it transforms an input image $\mathcal{P}^o$ into an output $\mathcal{P}^r$ via the mapping function $\mathcal{P}^r = f_\mu(\mathcal{P}^o)$. Network learning adjusts the parameters μ through minimizing three loss functions $\mathcal{L}_d$, $\mathcal{L}_p$, and $\mathcal{L}_{cc}$. Each loss function computes a scalar value $\mathcal{L}(.)$ measuring the difference between the input image and the generated one corresponding to three loss functions. The network is trained to minimize the weighted combination of the loss function:

$$\mu^* = \arg\min_\mu E_{\mathcal{V}}\left[\lambda_d\mathcal{L}_d + \lambda_{ca}\mathcal{L}_{ca} + \lambda_{cc}\mathcal{L}_{cc}\right]. \tag{14}$$

The parameters λ_d, λ_p and λ_c are all set to 1 in our training. In the following, we delve in detail the configuration of each loss function, apart of $\mathcal{L}_d$ defined in equation (9).

Findings in Gatys et al. [9] pointed out that the first few layers of VGG model [42] are sensitive to style while the later ones are sensitive to the content. Inspired by this finding, we feed $\mathcal{P}^o$ and $\mathcal{P}^r$ to the pre-trained model VGG and obtain the corresponding feature maps along layers $\Phi_i(\mathcal{P}^o)$ and $\Phi_i(\mathcal{P}^r)$, here $i = 1, \ldots, 5$. We then use them in the formulation of loss functions. Each layer in the encoder and decoder is injected into a *dual-MHA* and *dual-Conv3*, respectively.

Color consistency loss ($\mathcal{L}_{cc}$). Color consistency is an essential aspect in our results. With this loss function, we wish to penalize differences in colors, textures, common patterns, *etc.,* between the masked area and the entire image. We formulate this loss as:

$$\mathcal{L}_{cc} = \sum_{i=1,2} \frac{1}{C_i H_i W_i} \parallel \mathbf{G}(\Phi_i(\mathcal{P}^o)) - \mathbf{G}(\Phi_i(\mathcal{P}^r) * W_i T(t)) \parallel_2^2, \tag{15}$$

with

$$\mathbf{G}(\Phi_i(.)) = [\Phi_i(.)][\Phi_i(.)]^T. \tag{16}$$

The Gram matrix $\mathbf{G}$ captures the correlations between different feature maps, serving as a proxy for the image's textural patterns. The loss is the distance between the Gram matrices of the input and the target, also scaled dynamically with the time step t. This loss emphasizes reproducing the style or texture of the generated image $\mathcal{P}^r$.

Content alignment loss ($\mathcal{L}_{ca}$). To further enhance the alignment with the content of input images, we use feature maps extracted to compute the $\mathcal{L}_{ca}$. This loss is to measure how well the

generated image $\mathcal{P}^r$ intact with the source image $\mathcal{P}^o$. Rather than encouraging the pixels of $\mathcal{P}^r$ to exactly match the pixels of $\mathcal{P}^o$, we instead encourage them to have similar feature representation. Using pixel-wise comparisons could be computationally expensive and less efficient as they involve comparing every single pixel. Using features is more robust to such minor variations because they focus on higher-level features like edges, textures, and patterns. This helps in preserving the overall structure and details of the image. Formally, it is expressed as:

$$\mathcal{L}_{ca} = \sum_{j=1}^{4} \frac{1}{C_j H_j W_j} \parallel \Phi_j(\mathcal{P}^o) - \Phi_j(\mathcal{P}^r) * W_j T(t) \parallel_2^2 . \tag{17}$$

In this loss function, $j = 1 \ldots 4$ refers to the feature maps of the first four layers of VGG. With this function, we aim to make the network capture perceptual and texture differences between images that traditional pixel-wise loss functions (like MSE) might not capture.

4 EXPERIMENTAL RESULTS

4.1 Implementation Details

We implement our image reconstruction model in Pytorch [34]. All experiments are performed on a PC equipped with Intel Core i7-770 CPU, 16GB RAM and a GeForce RTX 2080 Ti GPU. The diffusion process is configured with $T = 1000$ timesteps using a linear variance schedule. During inference, we adopt 100 sampling steps for efficiency while preserving reconstruction quality. The model is trained with batch size 2 and the model is converged after 250 epochs. The training takes approximately 18 hours. For model optimizing, we use Adam optimizer with a learning rate of 1×10^{-3}. After training, our model spends approximately 50 seconds to inpaint an input PMA in size of 625 x 625.

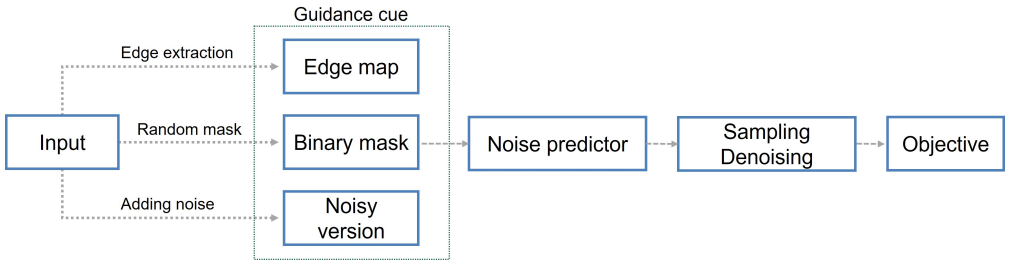


Fig. 4. Our training pipeline

In terms of training data, as stated earlier, we construct a self-collected dataset comprising 10.000 city map images retrieved from publicly accessible online sources using search queries of the form *"city map + [city name]"*, with city names were sampled from diverse geographical regions to ensure content variability. Retrieved images were filtered to retain RGB-format map-style visualizations with minimum resolution of 512×512 pixels, while excluding satellite imagery, 3D renderings, and images containing large watermarks or excessive textual overlays. All images were resized to 512×512 pixels and normalized to the range $[-1, 1]$ for diffusion training. For each image $\mathcal{I}^o$, a rectangle mask $\mathcal{M}$ was randomly generated with area ratio sampled from 15% to 40% of the image size and uniformly placed at random spatial locations to simulate diverse occlusions. The masked input is defined as $\mathcal{I}^{masked} = \mathcal{I}^o \odot \mathcal{M}$, where $\odot$ denotes element-wise multiplication. The dataset was randomly split into 8000 training, 1000 validation, and 1000 test images. The training pipeline of our framework is presented in Fig. 4.

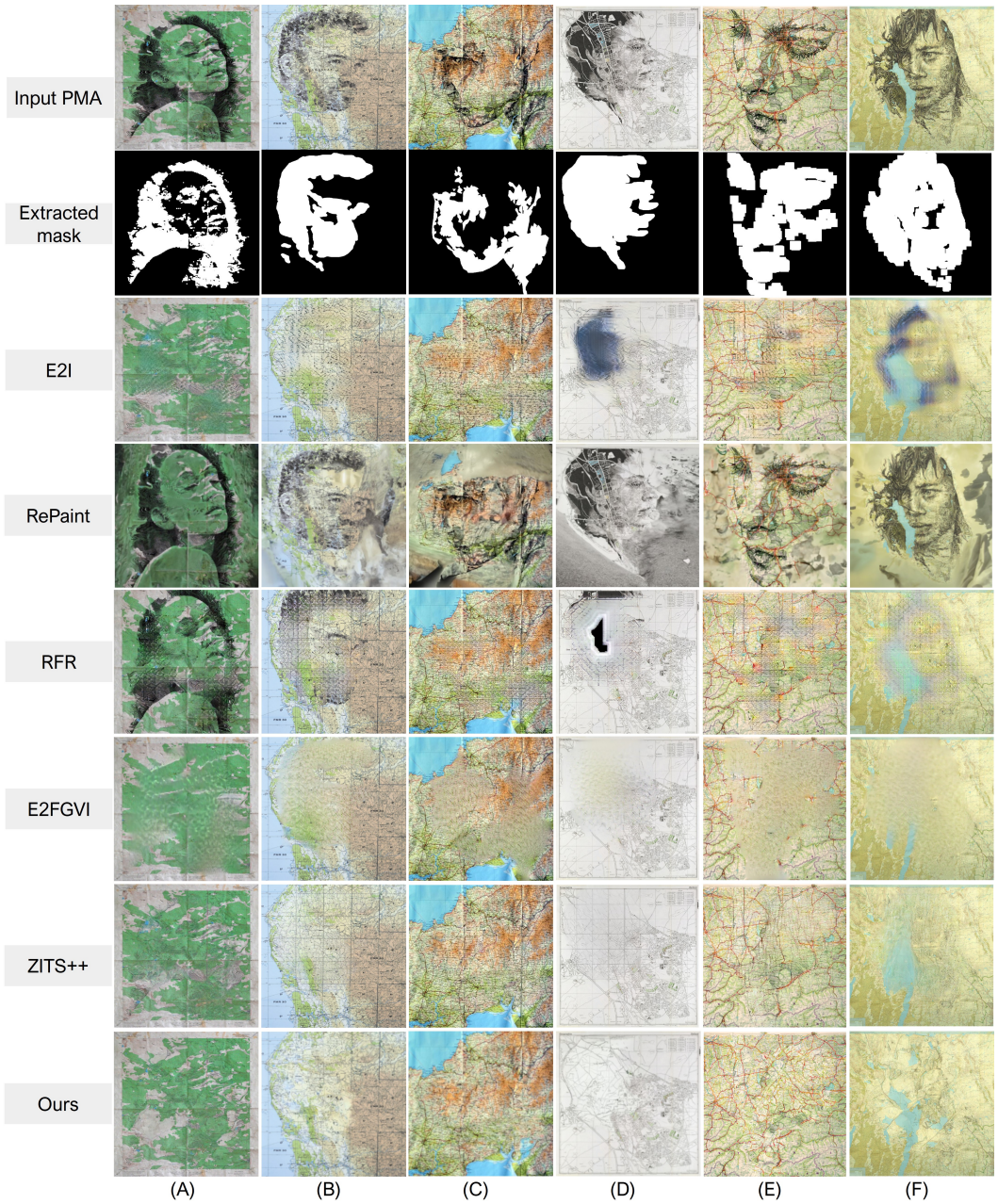


Fig. 5. Comparisons on various PMA images. From top to bottom: input PMA, our extracted $\mathcal{M}^e$, E2I [56], RePaint [31], RFR [22], E2FGVI [25], ZITS^{++} [4] and ours.

4.2 Results and Visual Comparisons

To demonstrate the effectiveness of our method, we tested it on various PMA contents. Fig.5 exhibits some typical results in our experiments, where the outputs of prior methods are arranged vertically

for comparison. Readers are encouraged to visit our project website[†] to access more visual results. The aspects that enable our results and model to advance prior work could be summarized as follows.

Our method demonstrates robustness in handling diverse PMA contents, including plain maps and intricate, colorful edges. This robustness is the result of our noise predictor design. By configuring *dual-Conv3* in the encoder and the *dual-MHA* in the decoder to the noise predictor, we can enhance our system's ability to understand high-level features of real-word map-based images. As shown in Fig.5, despite the diverse input PMA images, the generated results maintain a relatively equal visual effect. This underscores the generalization and stability of our method.

We can preserve the color consistency of the inner mask and entire image. This capability is beneficial from the utilization of distribution of $\mathcal{P}^o$ in the guidance cue both forward and backward path of the diffusion process. We can see this performance in most of the cases in Fig. 5, particularly samples (A), (B), (C), (E), and (F). The color intensity in these cases is relatively high, but we do not suffer from erase-like phenomenon. Besides, with the aid from color loss $\mathcal{L}_c$, color of the entire generated image is distributed evenly serving result harmonization and integrity with the source. Let us take samples (D) and (E), and we can see that our model can form a connecting edge with the color of these consistent with the surroundings.

For visual comparisons, Fig.5 presents the competition between our results and five state-of-the-art image inpainting methodologies including E2I [56], RePaint [31], RFR [22], E2FGVI [25], and ZITS^{++} [4]. From the results perspective, all the approaches are quite distinct, and the strong sense of harmonization in color and structure would be the viewer's first impression of our results. As we can see, inpainting on PMA data could be challenging to RePaint. The image content are damaged seriously in most of the testing samples. E2I and RFR are likely potential, although their results are not plausible. In some cases, such as (A), (B), and (C), E2I performs comparably to ours. However, the results of this comparison show that E2I and RFR share a mutual difficulty in preserving the coherent edge and color distribution. E2FGVI often produces over-smoothed textures and incomplete terrain structures, particularly in regions with dense geographic patterns. ZITS^{++} achieves competitive performance and preserves map structures more effectively than others; however, some cases still exhibit slight structural distortions or texture inconsistencies (*e.g.*, (D), (E), (F)). In comparison, our model produces more stable reconstructions with clearer road layouts and smoother terrain transitions. By incorporating the VOE-based mask and guidance-conditioned diffusion process, our method better preserves spatial continuity and reduces structural artifacts in the reconstructed regions.

4.3 Evaluations

4.3.1 VOE Mask evaluation. To quantitatively evaluate the VOE-generated masks, we manually annotate portrait regions for a subset of 100 randomly selected images which are from test split to avoid training bias. We compute the Intersection-over-Union (IoU) between the fine mask $\mathcal{M}^F$ and the ground-truth annotations. The average IoU is 0.82, indicating that the VOE module reliably localizes the portrait region with high spatial accuracy. After dilation with optimal κ, the coarse mask $\mathcal{M}^c$ achieves an average IoU of 0.88 with respect to the manually expanded boundary annotations.

To further analyze the influence of the dilation kernel κ on reconstruction performance, we vary κ from 0 to 40 and evaluate the reconstructed results using PSNR [12], SSIM [52], and LPIPS [66] on the validation set. Quantitative results are summarized in Table 1. As observed in the table, small

[†]http://graphics.csie.ncku.edu.tw/diffpmart

kernel sizes ($\kappa \leq 10$) result in lower PSNR, SSIM and higher LPIPS, indicating insufficient contextual coverage around the portrait boundary. When κ increases, reconstruction quality improves consistently and reaches optimal performance at $\kappa = 30$, which achieves the highest PSNR, SSIM and lowest LPIPS. Increasing κ beyond this point slightly degrades PSNR, meaning that excessive dilation introduces redundant context that weakens structural precision. These results confirm that an appropriately sized buffer region is essential for balancing boundary continuity and contextual guidance.

Table 1. Sensitivity analysis on dilation kernel size κ.

κ	PSNR ↑	SSIM ↑	LPIPS ↓
0	23.41	0.812	0.164
10	24.08	0.826	0.151
20	25.02	0.842	0.139
30	**26.31**	**0.861**	**0.121**
40	25.87	0.854	0.129

4.3.2 Results evaluations. To evaluate performance of our model, we evaluate it on two patches of samples: artist-created PMA (*i.e.,* without ground truth) and AI-created PMA (*i.e.,* with ground truth). For a thorough evaluation, we evaluate our results of these two patches with different metrics. Specifically, on the samples of artist-created PMA, we measure the harmonization of results with three aspects: (1) the color consistency, (2) edge coherence and continuity of the inpainted region versus its surrounding, and (3) the overall visual similarity between the input PMA and its result. On the patch of AI-created PMA, we adopt Le et al. [19], which takes a pair of plain map and portrait to produce PMAs. In this scenarios, we have the authentic structure to infer the structure of our reconstructed images. Therefore, we evaluate this patch on the similarity of the generated samples versus the ground-truth in terms of structure. Lastly, we conduct a cross-judgment on data from both patches via human visual perception.

C.1. Color Consistency Evaluation (CCE)

This metric assesses whether the color distribution of the reconstructed areas matches the surrounding. We use this metric to evaluate the color harmony between reconstructed features and their surroundings, which is essential for the realistic integration of different map elements. Inspired by the Earth Mover's Distance (EMD) [36], also known as the Wasserstein distance, a measure of the distance between two probability distributions over a region of space. We leverage EMD to quantify how similar the color distributions are, considering the entire color spectrum present in the images.

Let us denote $\mathcal{R}_m$ is the inpainted region, which is the subset of $\mathcal{P}^r$ where $\mathcal{M}^c$ is 1. And $\mathcal{R}_s$ is the surrounding region of $\mathcal{R}_m$, which is the subset of $\mathcal{P}^r$ where $\mathcal{M}^c$ is 0. We first extract the color histogram for both $\mathcal{R}_m$ and $\mathcal{R}_s$, yielding P and Q, respectively. This involves counting how frequently each color appears in both regions. Mathematically, the level of color consistency of $\mathcal{R}_m$ over $\mathcal{R}_s$ is expressed as:

$$\mathrm{CCE}(\mathcal{R}_m, \mathcal{R}_s) = \min \sum_{i=1}^{n} \sum_{j=1}^{m} f_{ij} d_{ij}. \tag{18}$$

Here f_{ij} represents the flow of distribution weight from bin i in P to bin j in Q; and d_{ij} is the distance between bin i to bin j; n and m are the number of bins in histograms P and Q, respectively. For this metric, the lower score represents a better color consistency.

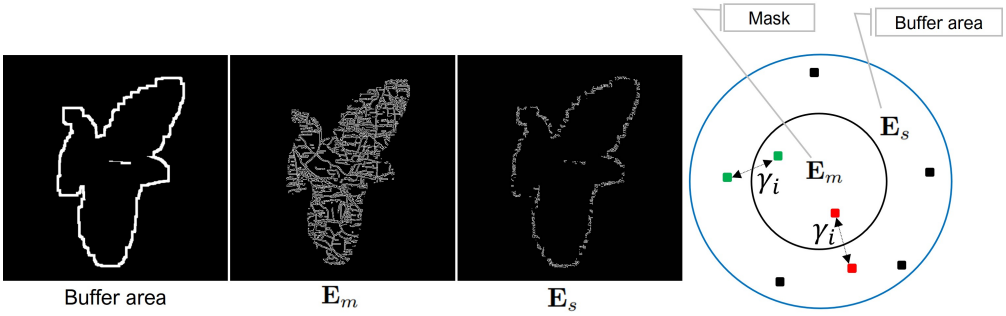


Fig. 6. Visualization of ECE

C.2. Edge Coherence and Continuity Evaluation (ECE)

The alignment and smooth transition of edges in resulting image are an important aspect in plausible reconstructed images. As ground truth is unavailable, we devise ECE metric to evaluate how seamlessly the edges within the inpainted region integrate with the edges in the surrounding areas of the image. To compute the alignment of edges between the two mentioned regions, we aim to quantify how well these edges align in terms of their spatial locations. A common approach is to measure the average distance from each edge pixel in the inpainted region to the nearest edge pixel in the surrounding area. For this computation, we elaborate as follows.

We first generate a so-called edge map ($\mathbf{E}$) by applying an edge detection algorithm to the generated image $\mathcal{P}^r$. In our experiments, we use the Candy edge [3] configured by kernel size 7×7, low threshold 10, and high threshold 200. This configuration is set to all of the samples in this evaluating experiment to ensure the fairness and stability of ECE score. Other alternatives for detecting edge may result in equivalent effect. From the edge map, we identify two sets of edge pixels $\mathbf{E}_m = \{p_{ij} \in \mathbf{E}\ s.t.,\ \mathcal{M}_{ij} = 1\}$ and $\mathbf{E}_s = \{p_{ij} \in \mathbf{E}\ s.t.,\ \mathcal{M}_{ij} = 0\}$, as visualized in Fig. 6. We then use the distance transform on the inverted edge map of $\mathbf{E}_s$ to get the distance of each pixel in $\mathbf{E}_s$ to the nearest non-edge pixel. Thereafter, inverting the logic to interpret these distances in the context of the nearest edge pixel.

For each edge pixel in $\mathbf{E}_m$, finding its nearest neighboring element in $\mathbf{E}_s$. The connectivity of a certain $p_i \in \mathbf{E}_m$ is defined by:

$$\gamma_i = \min\left(d(p_i^m, p_j^s)\right), \tag{19}$$

here $d(.)$ is the euclidean distance for spatial location of edge pixels $p_i^m \in \mathbf{E}_m$ and $p_i^s \in \mathbf{E}_s$. Accordingly, the level of continuity and edge coherence between the inpainted region and its surrounding is factorized as:

$$\mathrm{ECE}(\mathbf{E}_m, \mathbf{E}_s) = \frac{1}{N}\sum_{i=0}^{N} \gamma_i. \tag{20}$$

This equation computes the average of minimum distances γ_i to quantify the overall alignment. A lower ECE indicates better alignment, suggesting that the inpainted edges closely match the surrounding edges in position.

C.3. *Overall Visual Similarity Evaluation* (VSE)

To measure the overall visual similarity, we leverage the LPIPS (Learned Perceptual Image Patch Similarity) metric, proposed by Zhang et al. [66]. For this metric, we take the pair of images $\mathcal{P}^o$ and $\mathcal{P}^r$ as input. Each of image is first fed to a backbone $\mathcal{B}$ for feature extraction. Consequently,

Table 2. Analysis on Artist-created PMA samples

Metrics/Methods		Sample A	Sample B	Sample C	Sample D	Sample E	Sample F	Average
CCE ↓	E2I	0.71	0.68	0.84	0.51	0.68	0.77	0.7
	RFR	0.87	0.78	**0.82**	0.84	0.79	0.76	0.91
	Ours	**0.52**	**0.57**	0.86	**0.47**	**0.52**	**0.59**	**0.57**
ECE ↓	E2I	**0.18**	0.19	**0.21**	0.3	0.51	**0.53**	0.32
	RFR	0.27	0.23	0.22	053	0.76	0.75	0.46
	Ours	0.21	**0.16**	0.26	**0.27**	**0.32**	0.59	**0.3**
VSE ↓	E2I	0.29	**0.38**	0.2	0.31	0.27	0.37	0.3
	RFR	0.57	0.51	**0.19**	0.25	0.32	0.25	0.35
	Ours	**0.28**	**0.31**	0.49	**0.19**	**0.21**	**0.27**	**0.29**

the distance between feature at a certain layer k of two images is formulated as:

$$d_k = \frac{1}{H_k W_k} \parallel \mathcal{B}(\Phi_k(\mathcal{P}^o)) - \mathcal{B}(\Phi_k(\mathcal{P}^r)) \parallel_2^2 \tag{21}$$

For the backbone $\mathcal{B}$, we opt the pre-trained model ResNet [45] because this model is often used as a backbone for various tasks in satellite image analysis. Therefore, its strong feature representation capabilities is more appropriate to our current data than other alternatives.

To account for the perceptual relevance of different layers, we apply learned weights to the distances calculated at each layer. We note here that we do not hope the two images are identical since $\mathcal{P}^r$ is such an inpainted version. Thus, this metric is a relative measurement for the visual similarity, and we use it as an additional reference apart of CCE and ECE. The overall visual similarity of $\mathcal{P}^r$ over $\mathcal{P}^o$ is then expressed as:

$$\text{VSE}(\mathcal{P}^r, \mathcal{P}^o) = \sum_{k=1}^{L} w_k d_k, \tag{22}$$

where L is the number of layers in $\mathcal{B}$; w_k is the learned weight for layer k, indicating its importance in the perceptual similarity measurement. For this metric, the lower score implies better visual similarity.

The analysis on artist-created PMA is shown in table. 2. The artist-created PMA set contain six representative samples used for detailed evaluation. These samples correspond to the visual comparisons presented in Fig.5 and are selected to cover diverse portrait-map compositions. These samples cover diverse portrait-map compositions and are used for detailed case-oriented evaluation and complementary metric analysis on real-world PMA artworks. It can be seen that our method absolutely wins other competitors in terms of color consistency and edge continuity. For the overall visual similarity, high score is distributed on results of three methods. However, inspecting on value of single case, the difference of VSE on each case and average are not significant.

On the patch of AI-created PMA samples, as we have the ground truth map of AI-created PMAs, we measure quality of the results via different set of metrics. Let denote the authentic input map as $\mathcal{A}$ and the AI-created PMA as $\mathcal{P}^a$, the system of Le et al. [19] produces results by $\mathcal{A}$+portrait $\rightarrow \mathcal{P}^a$ protocol. As $\mathcal{P}^a$ is a stylized version of $\mathcal{A}$, instead of keeping accurate structure, we try to maintain

Table 3. Analysis on AI-created PMA samples

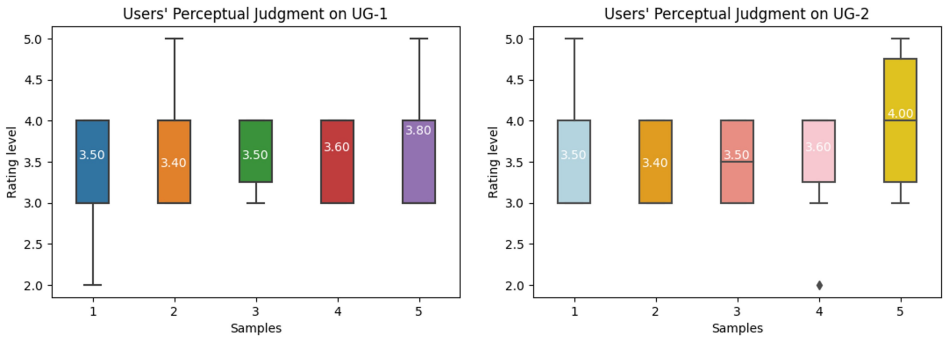
Metrics	Federal-face1	Federal-face2	Federal-face3	European-face1	European-face2	European-face3	Average
SSIM ↑	0.872	0.787	0.742	0.716	0.815	0.721	0.775
FID ↓	43.56	46.25	43.18	45.21	43.97	43.16	44.22
ECE ↓	0.23	0.318	0.272	0.341	0.297	0.315	0.29

the structural consistency between $\mathcal{A}$ and $\mathcal{P}^r$. Therefore, on this patch, we use SSIM [12], FID [59], and the above mentioned VSE to measure the similarity between $\mathcal{A}$ and $\mathcal{P}^r$. The SSIM is an index measuring the structural similarity between two images. When two images are nearly identical, their SSIM is close to 1. FID provides a symmetric measure of the distance between two distributions in the feature space of Inception-V3. Accordingly, the arguments of these metrics are $\mathcal{A}$ and $\mathcal{P}^r$. In patch, we evaluate on six samples, which are created by different maps, portraits, styles. Readers can access the project website for visual results of these samples. The analysis on this patch is presented in Table 3. We note here that methods E2I and RFR are absented from this experiment as the visual results in Fig.5 and analysis in Table 2 is obvious enough to demonstrate their performance on PMA data. The numerical values in this table demonstrate that the edge structure on our results are not at such a very good score but plausible enough. That is because we are inpainting $\mathcal{P}^a$, and $\mathcal{P}^a$ by itself is not in the same structure with $\mathcal{A}$ as stylization definitely damages image structure. Therefore, the scores on samples and average analysis in this table reveal we can produce image with structures asymptotic with authentic one.

To further access the validity of the proposed CCE and ECE metrics, we analyze their correlation with LPIPS [66] on the artist-created PMA set. We observe the Pearson correlation coefficients of 0.78 and 0.73 between LPIPS and CCE/ECE, respectively, indicating that lower CCE and ECE scores consistently correspond to lower LPIPS values, indicating improved perceptual realism. While LPIPS measure global perceptual similarity, CCE and ECE specifically quantify color harmony and boundary coherence, which are critical for map-portrait integration. Therefore, the proposed metrics complement rather than replacing standard perceptual measures.

E.4. *User study*

In this user study session, we use ten images, of which five are randomly selected from each of above patches. Ultimately, human judgment is critical to evaluating harmonization, as humans are particularly good at detecting visual and semantic anomalies. As we aim to assess the reality of our reconstructed results based on user judgment, we conduct a cross-user study. A total of 24 participants were invited and randomly divided into two groups (UG-1 and UG-2). On UG-1, participants were shown pairs of input PMA images and their reconstructed results (without ground truth reference) and asked to rate semantic plausibility on a 5-point Likert scale (higher is better). For UG-2, participants were shown input PMA, reconstructed results, and ground truth map, and asked to rate structural preservation relative to the ground truth. The results are presented in Fig. 7. The average score for UG-1 exceeds 3.5/5.0, indicating that reconstructed images are generally perceived as semantically coherent. For UG-2, the majority of votes is distributed in range of 3 to 4, with the average score exceeds 3.4/5.0, suggesting stable structural consistency across samples. As summarized in Fig.7, the rating distributions across participants exhibit relatively compact interquartile ranges and consistent median values, suggesting stable agreement among users. The absence of large dispersion indicates that user judgments are not dominated by outliners.



(a) (b)

Fig. 7. Analysis of user study on groups UG-1 (a) and UG-2 (b).

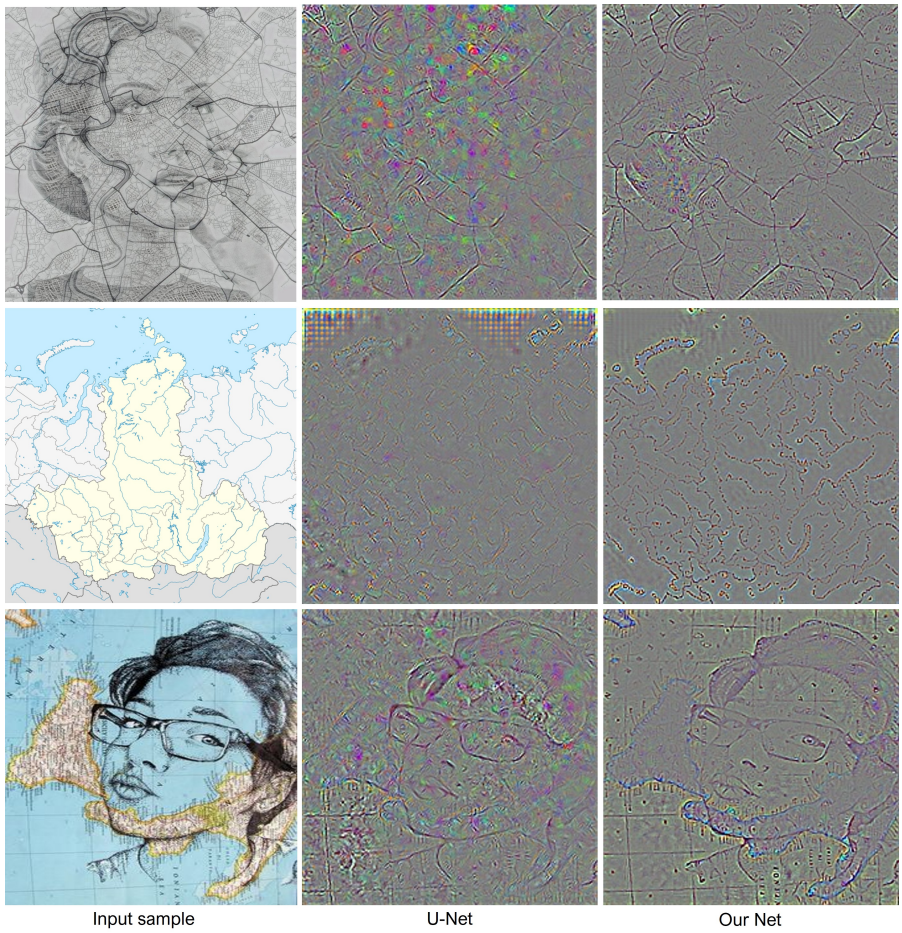


Fig. 8. Visualization of our tweaks versus the standard U-Net.

4.4 Ablation study

We now verify our design, which highlights our advancement through ablation studies. Fig.8 visualizes the effectiveness of our tweaks for the noise predictor, which is structured as a variation from a plain U-Net design. We test the activation of our design against the U-Net on three samples, including AI-created PMA, artist-created PMA, and a plain map to challenge the two models. The results show that our design can highlight the complex structure of various data, thus enabling our system to play with more general PMAs without suffering from abruption in edge connectivity.

The performance of our system is significantly improved by the utilization of guidance cue $\mathcal{S}$ during the diffusion phase. We remove $\mathcal{S}$ from both forward and backward processes to verify this configuration. That is, eliminating $\mathcal{S}$ from equations (7) to (9). The ablated results are presented in Fig. 9. It is obvious that without the guidance cue, it is challenging to obtain ideal inpainting results. We specify this phenomenon via two typical samples: color-focus on the first row and dense-edge-focus on the second. The mutual effect is that the color distribution is relatively weak without the guidance cue. The color in the entire image is not threaded to the masked region. This results in the phenomenon of this region being erase-like. In scenarios where the input PMA has a dense edge (second row), the network seems to struggle to connect edge attributes. Regarding loss functions, without $\mathcal{L}_{cc}$ could result in decent structure but the color challenging uncontrollable. Meanwhile, the absence of $\mathcal{L}_{ca}$ makes the results not realistic with variations of artifacts (see Fig.9(d)) not only texture but also reconstruction performance. The complete configuration serves our model, producing better visually pleasing results compared to ablated ones.

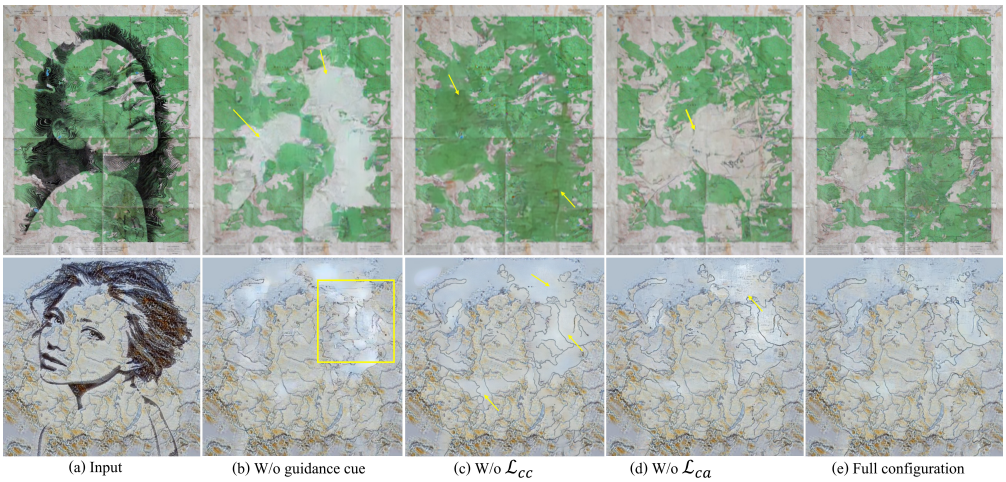


Fig. 9. Ablated results in our designed model.

The extracted mask plays an essential role in balancing the noise distribution in the diffusion process's forward path. Therefore, we verify the effect of mask area on the quality of results. It could be seen via the visualizations in Fig.10. Obviously, using fine-mask $\mathcal{M}^F$ alone (*i.e.,* $\kappa = 0$) is insufficient for constructing the region. Increasing the kernel κ in equation (2) can work, but it still causes artifacts, making the result unpleasing (see (b)). If the kernel κ is too small, $\mathcal{M}^c$ is too sharp to be identical to $\mathcal{M}^F$. Therefore, it is not sufficient to provide color information. In (b), $\kappa = 10$ is not big enough to cover the portrait region. Thus, $\mathcal{M}^c \odot \mathcal{P}^o$ still contains information about the

portrait, particularly the color/pattern. As a result, this information causes pattern noise in the generated image. And with appropriate $\kappa = 30$, we obtain decent reconstructed image.

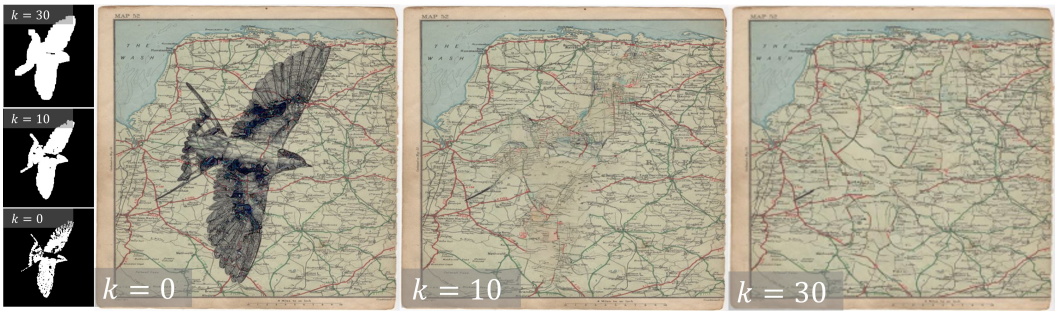


Fig. 10. Ablated results of extracted mask.

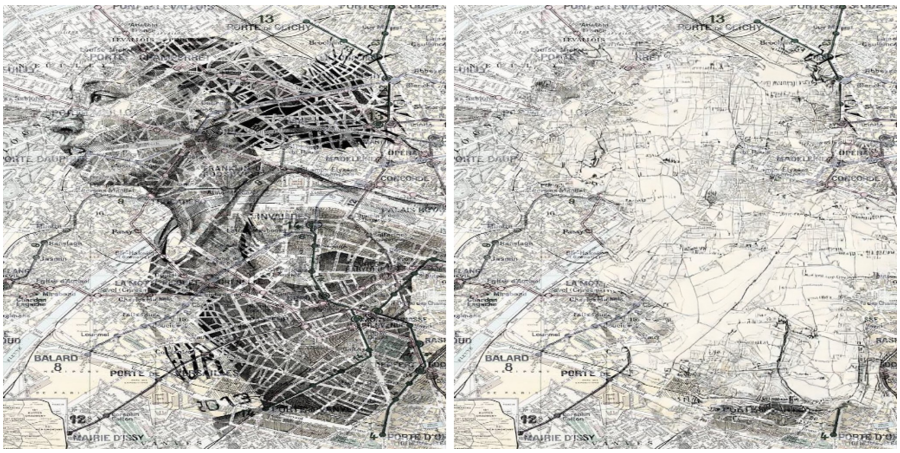


Fig. 11. A sample of our bad case.

4.5 Limitation

Although our model can perform well and be stable on intensive PMA samples, it also has certain limitations. The failure phenomena are probably visualized via the typical sample in Fig. 11. When portraits exhibit ambiguous boundaries, strong texture overlap with map structures, or highly stylized edge patterns, the VOE mask may not precisely capture semantic boundaries, leading to boundary artifacts during diffusion reconstruction. If the masked region occupies a substantial portion of the image (*e.g.*, over 50%), insufficient contextual information may result in over-smoothed textures or structural inconsistencies. Besides, maps with inconsistent color schemes or compression artifacts can propagate irregular patterns into reconstructed regions, reducing color harmony and edge coherence. In terms of generalization, the Diff-PMArt is specifically designed for map-portrait artworks, where structured geographic layouts provide strong spatial priors. The current formulation does not explicitly incorporate domain-adaptive mechanism; therefore, direct application to other visual domains (*e.g.*, medical imaging or abstract artistic forms) may not yield decent performance without architectural modification or retraining on domain-specific data. We leave these directions for our future works.

5 CONCLUSIONS

In this work, we present Diff-PMArt, a novel framework for the unsupervised inpainting of Portrait-Map-Art (PMA) images, which addresses the unique challenge of reconstructing complex map backgrounds after the removal of integrated portraits. Unlike prior approaches that rely on predefined rectangular masks and labeled data for training, our system demonstrates the ability to handle unknown regions without annotations, offering a more flexible and intractable solution for image inpainting. Our approach not only overcomes the limitations of traditional methods but also significantly improves the ability to generate realistic and visually coherent results in complex structured images. We expect Diff-PMArt is not only a significant step forward for structured image inpainting but also holds promise for practical applications such as artistic restoration, heritage preservation, and interactive art. With the growing demand for sophisticated image editing tools across social media and other digital platforms, our framework provides an important solution for high-quality, accurate image reconstruction in complex domains. In future work, we aim to further refine our framework by exploring more advanced object extraction techniques and expanding the generalizability of our model to other structured image-type applications. Additionally, integrating user-guided inpainting mechanisms or incorporating interactive features could potentially enable Diff-PMArt to be utilized in more interactive art and restoration applications.

ACKNOWLEDGEMENT

This work is supported by the National Science and Technology Council, Taiwan under Grant 113-2221-E006-161-MY3 and Grant 114-2221-E-006-114-MY3.

REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* 28, 3 (2009), 24.

[2] Nian Cai, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. 2017. Blind inpainting using the fully convolutional neural network. *The Visual Computer* 33 (2017), 249–261.

[3] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

[4] Chenjie Cao, Qiaole Dong, and Yanwei Fu. 2023. ZITS++: Image inpainting by improving the incremental transformer on structural priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12667–12684.

[5] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. Multi-head attention: Collaborate instead of concatenate. *arXiv preprint arXiv:2006.16362* (2020).

[6] Tausif Diwan, G Anirudh, and Jitendra V Tembhurne. 2023. Object detection using YOLO: Challenges, architectural successors, datasets and applications. *Multimedia Tools and Applications* 82, 6 (2023), 9243–9275.

[7] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. 2020. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 2777–2787.

[8] Lu Feihong, Chen Hang, Li Kang, Deng Qiliang, Zhao Jian, Zhang Kaipeng, and Han Hong. 2023. Toward high-quality face-mask occluded restoration. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 1 (2023), 1–23.

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2414–2423.

[10] James Hays and Alexei A Efros. 2007. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)* 26, 3 (2007), 4–es.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[12] Alain Hore and Djemel Ziou. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*. IEEE, 2366–2369.

[13] Cong Hu, Xiao-Zhong Wei, and Xiao-Jun Wu. 2024. DIRformer: A Novel Image Restoration Approach Based on U-shaped Transformer and Diffusion Models. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[14] Jingjing Hu, Dan Guo, Kun Li, Zhan Si, Xun Yang, Xiaojun Chang, and Meng Wang. 2025. Unified static and dynamic network: efficient temporal filtering for video grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).

[15] Nisha Huang, Yuxin Zhang, Fan Tang, Chongyang Ma, Haibin Huang, Weiming Dong, and Changsheng Xu. 2024. Diffstyler: Controllable dual diffusion for text-driven image stylization. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

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

[17] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1646–1654.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[19] Thi-Ngoc-Hanh Le, Ya-Hsuan Chen, and Tong-Yee Lee. 2023. Structure-aware Video Style Transfer with Map Art. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 3s (2023), 1–25.

[20] Thi-Ngoc-Hanh Le, Tong-Yee Lee, Shih-Syun Lin, and Weiming Dong. 2024. Deep learning-based importance map for content-aware media retargeting. *Multimedia Tools and Applications* (2024), 1–22.

[21] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing* 479 (2022), 47–59.

[22] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. 2020. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 7760–7768.

[23] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang. 2018. Single image dehazing via conditional generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8202–8211.

[24] Xin Li, Yulin Ren, Xin Jin, Cuiling Lan, Xingrui Wang, Wenjun Zeng, Xinchao Wang, and Zhibo Chen. 2023. Diffusion Models for Image Restoration and Enhancement–A Comprehensive Survey. *arXiv preprint arXiv:2308.09388* (2023).

[25] Zhen Li, Cheng-Ze Lu, Jianhua Qin, Chun-Le Guo, and Ming-Ming Cheng. 2022. Towards an end-to-end framework for flow-guided video inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 17562–17571.

[26] Pengwei Liang, Junjun Jiang, Xianming Liu, and Jiayi Ma. 2024. Image deblurring by exploring in-depth properties of transformer. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[27] Jiayu Lin and Yuan-Gen Wang. 2024. TSFormer: Tracking Structure Transformer for Image Inpainting. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 12 (2024), 1–23.

[28] Chang Liu, Shunxin Xu, Jialun Peng, Kaidong Zhang, and Dong Liu. 2024. Towards Interactive Image Inpainting via Robust Sketch Refinement. *IEEE Transactions on Multimedia* (2024).

[29] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*. 85–100.

[30] Zhisheng Lu, Juncheng Li, Hong Liu, Chaoyan Huang, Linlin Zhang, and Tieyong Zeng. 2022. Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 457–466.

[31] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 11461–11471.

[32] K Nazeri. 2019. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *arXiv preprint arXiv:1901.00212* (2019).

[33] Ozan Özdenizci and Robert Legenstein. 2023. Restoring vision in adverse weather conditions with patch-based denoising diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 10346–10357.

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).

[35] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2536–2544.

[36] Ofir Pele and Michael Werman. 2009. Fast and robust earth mover's distances. In *2009 IEEE 12th international conference on computer vision*. IEEE, 460–467.

[37] Shruti S Phutke, Ashutosh Kulkarni, Santosh Kumar Vipparthi, and Subrahmanyam Murala. 2023. Blind image inpainting via omni-dimensional gated attention and wavelet queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1251–1260.

[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern*

recognition. 7479–7489.

[39] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. 2022. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4713–4726.

[40] Cai Shang, Yu Zeng, Shu Yang, Xu Jia, Huchuan Lu, and You He. 2023. Deformable Dynamic Sampling and Dynamic Predictable Mask Mining for Image Inpainting. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[41] Chiao-Yin Shih, Ya-Hsuan Chen, and Tong-Yee Lee. 2021. Map art style transfer with multi-stage framework. *Multimedia Tools and Applications* 80 (2021), 4279–4293.

[42] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[43] Yang Song and Stefano Ermon. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* 32 (2019).

[44] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[45] Sasha Targ, Diogo Almeida, and Kevin Lyman. 2016. Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029* (2016).

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. Advances in neural information processing systems. *Advances in neural information processing systems* 30, 2017 (2017).

[47] Junke Wang, Shaoxiang Chen, Zuxuan Wu, and Yu-Gang Jiang. 2022. Ft-tdr: Frequency-guided transformer and top-down refinement network for blind face inpainting. *IEEE Transactions on Multimedia* 25 (2022), 2382–2392.

[48] Juan Wang, Chunfeng Yuan, Bing Li, Ying Deng, Weiming Hu, and Stephen Maybank. 2023. Self-prior guided pixel adversarial networks for blind image inpainting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 10 (2023), 12377–12393.

[49] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. 2022. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050* (2022).

[50] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*. 0–0.

[51] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. 2020. Vcnet: A robust approach to blind image inpainting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*. Springer, 752–768.

[52] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[53] Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhang, and Yanfeng Wang. 2023. Dr2: Diffusion-based robust degradation remover for blind face restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1704–1713.

[54] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. 2022. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16293–16303.

[55] Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. 2024. A Survey of Camouflaged Object Detection and Beyond. *arXiv preprint arXiv:2408.14562* (2024).

[56] Shunxin Xu, Dong Liu, and Zhiwei Xiong. 2020. E2I: Generative inpainting from edge to image. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 4 (2020), 1308–1322.

[57] Mingzhe Yang, Sihao Lin, Changlin Li, and Xiaojun Chang. 2025. Let LLM Tell What to Prune and How Much to Prune. In *International Conference on Machine Learning*. PMLR, 70833–70849.

[58] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 5505–5514.

[59] Yu Yu, Weibin Zhang, and Yun Deng. 2021. Frechet inception distance (fid) for evaluating gans. *China University of Mining Technology Beijing Graduate School* 3 (2021).

[60] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. 2019. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1486–1494.

[61] Dandan Zhan, Jiahao Wu, Xing Luo, and Zhi Jin. 2024. Learning from Text: A Multimodal Face Inpainting Network for Irregular Holes. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

[62] Cong Zhang, Wenxia Yang, Xin Li, and Huan Han. 2024. MMGInpainting: Multi-Modality Guided Image Inpainting Based On Diffusion Models. *IEEE Transactions on Multimedia* (2024).

[63] Jing Zhang and Dacheng Tao. 2019. FAMED-Net: A fast and accurate multi-scale end-to-end dehazing network. *IEEE Transactions on Image Processing* 29 (2019), 72–84.

[64] Jie Zhang and Wanming Zhai. 2022. Blind attention geometric restraint neural network for single image dynamic/defocus deblurring. *IEEE Transactions on Neural Networks and Learning Systems* 34, 11 (2022), 8404–8417.

[65] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. 2020. Deblurring by realistic blurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2737–2746.

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

[67] Yushu Zhang, Zhibin Fu, Shuren Qi, Mingfu Xue, Xiaochun Cao, and Yong Xiang. 2023. PS-Net: A Learning Strategy for Accurately Exposing the Professional Photoshop Inpainting. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[68] Yuxin Zhang, Fan Tang, Weiming Dong, Thi-Ngoc-Hanh Le, Changsheng Xu, and Tong-Yee Lee. 2023. portrait map Art generation by Asymmetric Image-to-Image translation. *Leonardo* 56, 1 (2023), 28–36.

[69] Yang Zhang, Wang Zhou, Gaoyuan Zhang, David Cox, and Shiyu Chang. 2022. An Adversarial Framework for Generating Unseen Images by Activation Maximization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3371–3379.

[70] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. 2021. RGB-D salient object detection: A survey. *Computational Visual Media* 7 (2021), 37–69.

[71] Yu Zhou, Zhihua Chen, Ping Li, Haitao Song, CL Philip Chen, and Bin Sheng. 2022. FSAD-Net: feedback spatial attention dehazing network. *IEEE Transactions on Neural Networks and Learning Systems* 34, 10 (2022), 7719–7733.