An approach for medical cancer image generation leveraging pretrained Stable Diffusion v2.1

Kien T. Pham
The Hong Kong University of Science and Technology
tkpham@connect.ust.hk

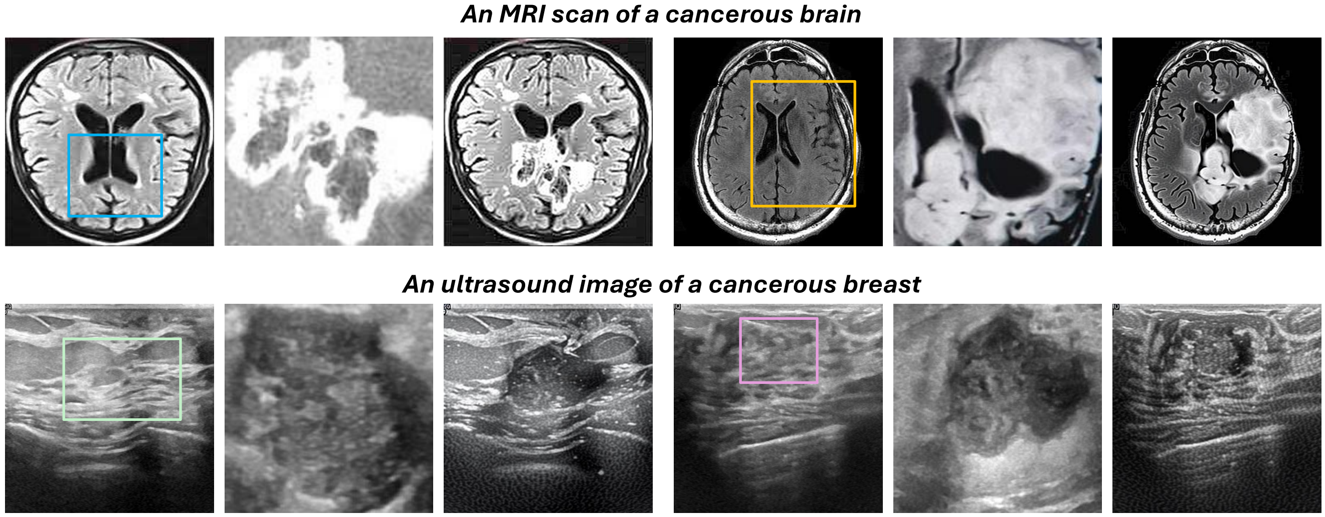
An MRI scan of a cancerous brain
An ultrasound image of a cancerous breast

Figure 1. Medical cancer image composition targets at seamlessly incorporating a foreground tumour object into a specific non-tumorous background image. Our training-free framework enables text-to-image diffusion probabilistic models with the controllability to effectively and efficiently conquer this task.

Abstract

Diffusion Probabilistic Models (DPMs) have emerged as a prominent research focus due to their exceptional generative capabilities in data synthesis. Despite the computational challenges posed by the iterative sampling process, DPMs are highly valued in the field of medical imaging, particularly for their ability to produce high-fidelity and diverse data. Such advantages have led to their widespread adoption in various medical imaging tasks such as cancer classification of pathology images, tumor detection and segmentation, or breast image registration... This report tackles an under-explored application of DPMs dubbed image composition for cancer image generation which aims at seamlessly incorporating a tumour (foreground) object into a specific non-tumorous (background) image. Due to domain disparity, the majority of works involving DPMs in medical imaging often requires training or fine-tuning DPMs on corresponding datasets which are expensive and time-consuming. In contrast, we directly leverage text-driven DPMs pretrained on generic domains to achieve our goal without any additional training. Our experiments demonstrate the strong potentiality and capabilities of the proposed method in generating highly realistic composited medical cancer images. Our proposed method offers a potent and efficient solution to the pervasive challenge of limited labeled datasets in medical imaging domains, mitigating the necessity for resource-intensive annotation processes.

1. Introduction

In general context, image composition involves incorporating distinct objects from disparate sources to create a cohesive image within a specified visual context, commonly referred to as image-guided composition. For example, consider the scenario of seamlessly integrating a fox into a pre-existing artwork, such as an oil or sketchy painting. The objective is to generate a composited image where the

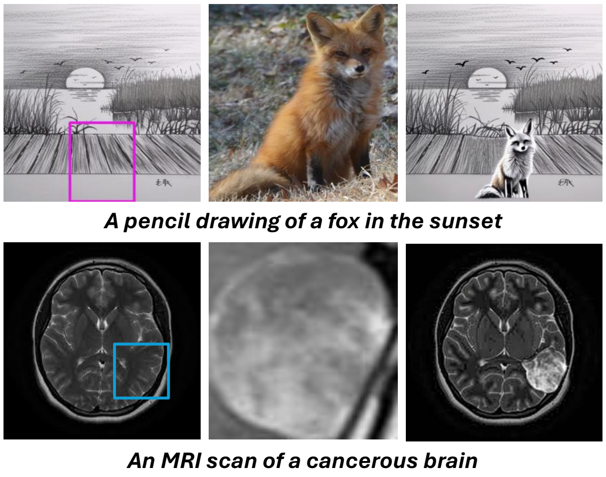
A pencil drawing of a fox in the sunset
An MRI scan of a cancerous brain

Figure 2. Image-guided composition in general context and in medical imaging domain.

fox appears harmoniously into the background while retaining its appearance and the background remains untouched, akin to the meticulous work of an artist (See Fig. 2). This task poses significant challenges, as it necessitates maintaining consistent illumination and retaining visual appearance. The complexity is additionally exacerbated in scenarios that the source images originate from diverse domains. Despite the remarkable progress of text-to-image models [7, 14, 45, 46, 48, 63] in text-guided image generation, the inherent ambiguous nature of natural language poses challenges in describing and conveying accurate and nuanced visually detailed inputs, even with highly meticulous text prompts. Personalized concept learning methods [18, 19, 27, 30, 47] effectively address this challenge but require computationally expensive instance-by-instance optimization and can only produce conceptualized images with specified backgrounds. Recent works [54, 62] have demonstrated the potential of DPMs for image-controlled composition by incorporating additional images as conditional guidance. Notwithstanding, those models are typically retrained from pre-trained DPMs on customized datasets, which can deteriorate the learned prior knowledge of the model. Consequently, they exhibit subpar compositional abilities beyond their training domain and necessitate intensive computational resources.

In the medical imaging context, image-guided composition in general downstream tasks and in cancer image generation for specific remain under-explored. One of the obvious reasons is the lack of labeled medical datasets necessary to train a decent network to achieve a certain goal while acquiring expert annotation is extremely expensive and time-consuming. Given the power of large text-to-image DPMs pre-trained on comprehensive language-vision datasets of generic domains, we ask a question: *how could these models be utilized for medical cancer image composition without the needs for costly training or fine-tuning, thus preserving their learned diverse priors?* To answer this question, we introduce a simple yet effective and efficient technique which equips text-to-image DPMs with the ability to conduct medical cancer image composition without necessitating any extra fine-tuning, training, additional data, or optimization, facilitating the generation of highly realistic composition outcomes. To our knowledge, the proposed approach is the first method to specifically handle medical-based image composition in training-free manner. The proposed framework is not only compatible with many DPM samplers but also allows for completed generation within 20-25 steps, directly leveraging rich semantic knowledge to enable effective image-guided compositions.

Our approach comprises of two stages which are image inversion to invert the input non-tumorous medical image (background) and a tumour image (foreground) into their corresponding latent representations, and composition to incorporate the obtained foreground and background latents in generating final seamless composited results. Firstly, image inversion, which aims at reconstructing an input image while allowing for a certain controllability, is a challenging yet crucial step for DPM-based image editing frameworks involving real images [11, 21, 29, 31, 40, 42, 43, 58]. The most commonly used approach is DDIM [52], shortened for denoising diffusion implicit models, inversion that has shown to be workable for unconditional DPMs. However, for text-conditioned DPMs and medical imaging, it falls short of expectation. To overcome this situation, we adopt a recently proposed method dubbed exceptional prompt inversion from [39] to precisely invert input images into their respective latent codes.

Secondly, upon obtaining accurate inverted latents, we then conduct reverse process to gradually denoise them while performing composition intertwine. Unlike [39] which starts composing at initial timestep when the noisy latents are still at chaotic stage and contain very limited information suitable for composition, we propose to delay the initiation until later timestep at semantic stage of denoising process. This not only allows for more useful features of background and foreground images being reconstructed then incorporated into composited outcomes, but also ensures enough time for the DPMs to rectify the composition hence produce highly realistic images.

Overall, our key contributions are as follow:

- We introduce the first training-free DPMs-based framework for medical cancer image composition that facilitates the generation of highly realistic cancer images from a tumour object and a non-tumorous background images.
- We propose a novel technique to initiate the composition process that guarantees desired composited outcomes.

- We showcase the potentiality and effectiveness of our proposed method through extensive experiments on various medical visual contexts.

2. Related Works

2.1. Image Composition

Even though image composition finds applications in diverse domains, including e-commerce, entertainment, and data augmentation for downstream tasks [15, 34], its adoption in medical imaging is an under-explored direction. Therefore, this section will mainly discuss relevant works in general contexts and domains. Image composition can be broadly classified into two primary directions which are text-guided and image-guided composition.

Text-guided Composition. Text-guided composition involves generating images based solely on a textual description, without conditioning on the visual appearance of input objects as long as the text prompts precisely align with their semantics [2, 3, 8, 17, 36]. Despite the advancements of text-driven models, they are susceptible to semantic ambiguities [17, 45], such as attribute leakage, missing objects, and attribute interchange, which can lead to resulting images deviating significantly from the user's intent [17, 45]. Consequently, extensive and meticulous prompt engineering [59] is usually required to obtain satisfactory outcomes.

Image-guided Composition. In contrast, image-guided composition encapsulates specific objects and scenarios from user-given images, optionally supplemented by an input text prompt. Notwithstanding, the inclusion of extra images introduces challenges, particularly in cases when merging images from disparate visual contexts or domains, which is a central focus of painterly image harmonization [6, 38, 64]. Traditionally, image-guided composition has been decomposed into sub-tasks [41], such as object placement [4, 9, 32, 57, 65], image blending [60, 64], image harmonization [10, 12, 24, 61], and shadow generation [23, 33, 50, 67], each typically addressed by distinct models and pipelines.

2.2. Medical Image Generation

Image composition, a sub-field of image generation, remains relatively under-exploited in the context of medical imaging. However, a growing body of research utilizes image generation for diverse downstream tasks within this domain. Notably, [28] fine-tunes Latent Diffusion Model (LDM) [46] using Dreambooth [47] for generation of chest and lung X-ray images, cancer brain magnetic resonance imaging (MRI) scans, and contrast enhanced spectral mammography (CESM) images, demonstrating the capabilities of DPMs in capturing attributes specific to oncology within various medical imaging modalities. [1] utilizes Deepfake to produce MRI brain scans which are then used to improve robustness of segmentation networks. [55] introduces StylePix2pix that allows for one-to-many lung cancer CT image generation with complex tumor shapes based on a free-form sketch. [5, 26] leverages Generative Adversarial Network (GAN) [20] to synthesize skin cancer images.

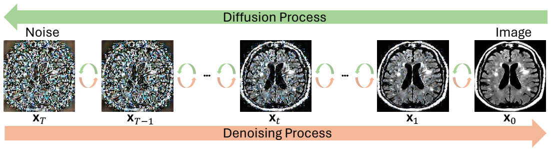

Figure 3. Diffusion and denoising processes in DPMs [16].

3. Preliminary

Diffusion probabilistic models (DPMs) represent a novel approach to generative modeling, employing a neural network to approximate the representation of a distribution of Gaussian noises $\epsilon_t \sim \mathcal{N}(\mu_t, \sigma_t)$. These noises possess the ability to perturb observed data $\mathbf{x}$, transforming it into noise that conforms to a standard normal distribution $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This transformation process is termed the "forward diffusion process". Subsequently, the models progressively recover the original data sample $\hat{\mathbf{x}}$ from the noise variable $\mathbf{z}$ via a procedure referred to as the "reverse denoising process" as illustrated in Fig. 3.

Forward Diffusion Process. [51] pioneered the development of DPMs, introducing the principle of transforming a known simple distribution into an unknown distribution through a Markov chain process. Formally, given a data distribution $q(\mathbf{x}_0)$, the forward diffusion process is constructed using a discrete-time Markov chain $\{\mathbf{x}_t | 0 \leq t \leq T, t \in \mathbb{N}\}$ and a transition probability $q(\mathbf{x}_t|\mathbf{x}_{t-1})$. Leveraging the intrinsic property of Markov chain, we can express the relationship between $q(\mathbf{x}_0)$ and the stationary distribution $q(\mathbf{x}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ as:

$$\begin{aligned} q(\mathbf{x}_1, \ldots, \mathbf{x}_T|\mathbf{x}_0) &= \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}), \\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \end{aligned} \tag{1}$$

in which $\beta_t \in (0, 1)$ is a predefined set of increasing diffusion hyperparameters, representing the magnitude of the noise induced to the initial signal. By simply reparameterizing $\alpha_t = 1 - \beta_t, \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and leveraging the transition probability $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, $\mathbf{x}_t$ can be directly derive from $\mathbf{x}_0 \sim p(\mathbf{x}_0)$ and noise $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t \tag{2}$$

Reverse Denoising Process. Initiating from $\mathbf{x}_T$, the reverse denoising process can be defined by a reverse-time Markov chain with a transition probability estimated by the conditional probability $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$, which is learnable under the following formulation:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \quad (3)$$

In this formulation, $\mu_\theta(\mathbf{x}_t, t)$ and $\Sigma_\theta(\mathbf{x}_t, t) = \sigma_t^2\mathbf{I}$ represent the mean and variance, respectively, as determined by a learned neural network $\epsilon_\theta(\mathbf{x}_t, t)$ with θ denoting its parameters. The value of σ_t is typically set to a fixed value β_t or $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}$. Utilizing this learned transition probability, we can commence the sampling process by randomly initializing $\mathbf{x}_T \sim p(\mathbf{x}_T)$. Subsequently, we iteratively sample $\mathbf{x}_{t-1}$ until reaching timestep $t = 0$, ultimately obtaining the generated result $\hat{\mathbf{x}}_0 \sim p(\mathbf{x}_0)$. The entire sampling procedure can be formally expressed as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t\mathbf{z}, \quad (4)$$

in which $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $t = T, \ldots, 1$.

To train the neural network $\epsilon_\theta(\mathbf{x}_t, t)$, [22] propose employing a simplified optimization objective, which involves minimizing the following training loss function:

$$\mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2\right] \quad (5)$$

The trained neural network $\epsilon_\theta(\mathbf{x}_t, t)$ can be conceptualized as a denoising model that estimates the noise ϵ_t added to $\mathbf{x}_t$ in Eq. 2. Consequently, the reverse process effectively corresponds to the progressive removal of Gaussian noise, ultimately yielding a clean image $\hat{\mathbf{x}}_0$.

4. Methodology

4.1. Problem Formulation

The objective for medical cancer image composition using text-driven DPMs is to use a tumour-free background image $\mathbf{I}_{bg}$, a tumour object image $\mathbf{I}_{obj}$ with segmentation mask $\mathbf{M}_{obj}$, a text prompt $\mathbf{P}$, and a user-inputted binary mask $\mathbf{M}_{user}$ which specifies the desired area of $\mathbf{I}_{bg}$ to incorporate the tumour $\mathbf{I}_{obj}$, to generate a tumorous image $\mathbf{I}_{res}$. The resulting image $\mathbf{I}_{res}$ should acquire these following properties: 1. $\mathbf{I}_{res}$ should faithfully preserve the identifying features of the tumour object within the designated mask, *i.e.* $\text{ID}(\mathbf{I}_{res} \odot \mathbf{M}_{user}) \approx \text{ID}(\mathbf{I}_{obj})$; 2. The background area outside mask should remain unchanged, *i.e.* $\mathbf{I}_{res} \odot (\mathbf{1} - \mathbf{M}_{user}) \approx \mathbf{I}_{bg} \odot (\mathbf{1} - \mathbf{M}_{user})$; 3. The transition from inside to outside the mask areas should be visually imperceptible and artifact-free. We propose a novel training-free framework that utilizes pre-trained LDM [46] for medical cancer image composition in image-guided manner. To our knowledge, this is the first approach specifically tackling this task in training-free manner, which can operate within 20 sampling steps and consists of two primary stages: image inversion (Section 4.2) and image composition generation (Section 4.3), as depicted in Fig. 5.

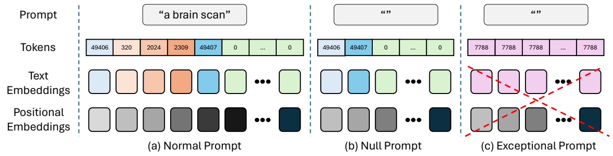

Figure 4. Illustrations for difference among three prompt embedding calculations: (a) Normal Prompt, (b) Null Prompt, and (c) Exceptional Prompt [39].

4.2. Image Inversion

Controllable manipulation on real images often requires an accurate image inversion process to identify the according latent code representation. This latent representation not only facilitates meaningful manipulation through editability but also enables faithful reconstruction of the inputted images [4, 56].

ODE Inversion. We employ the exceptional prompt inversion introduced in [39] as the inversion technique which is developed on top of an ODE solver named DPM-Solver++ [37]. Unlike many diffusion works for image editing [11, 21, 29, 31, 42, 58] which adopts DDIM [52] inversion to invert inputs into latent representations, [39] finds that it may result in sub-optimal inverted noises. Given that DDIM has been demonstrated to be a first-order discretization of the corresponding probability flow ordinary differential equation (ODE) [49, 52, 53], several samplers [25, 35] have been proposed for solving this diffusion ODE, especially for DPM-Solver++ [37] which achieves better alignment between the forward and backward ODE trajectories in the high-order, not only enabling rapid sampling within 10-20 steps but also yields better latent representations. Therefore, similar to [39], we leverage it to conduct all image inversions of DPMs.

Exceptional Prompt Formulation. Within the unconditional settings $\epsilon_\theta(\mathbf{x}_t, t)$ where no text prompt is provided, solving the diffusion ODE from 0 to T results in accurate latent representation $\mathbf{x}_T$ of the real inputted image $\mathbf{x}_0$. Yet, in the text-conditioned configurations with $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{P})$, prior approaches [4, 56, 58] have revealed that the inversion process is susceptible to noticeable reconstruction errors. This instability arises from the inherent limitations of classifier-free guidance (CFG) [13, 22] below:

$$\hat{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{P}, \varnothing) = s \times \epsilon_\theta(\mathbf{x}_t, t, \mathbf{P}) + (1-s) \times \epsilon_\theta(\mathbf{x}_t, t, \varnothing), \quad (6)$$

where $\mathbf{P}$ and $\varnothing$ denotes normal and null text prompts. This problem persists even without CFG that $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{P})$ and

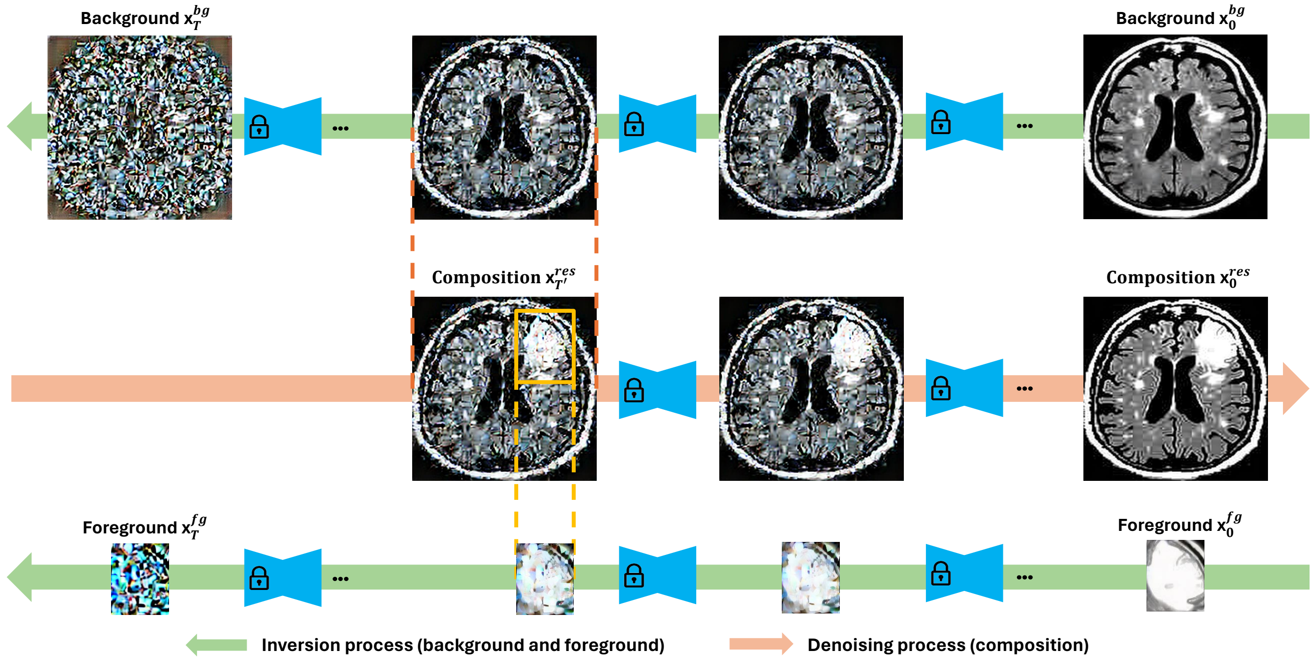

Figure 5. Illustration for the proposed training-free medical cancer image composition framework. Tumor-free background image $\mathbf{x}_0^{bg}$ and tumour object image $\mathbf{x}_0^{fg}$ will be inverted into their respective latent codes $\mathbf{x}_T^{bg}$ and $\mathbf{x}_0^{fg}$. Then for a selected timestep T', the composition process is initiated by incorporating the obtained latent codes. Finally, the rest of the composing steps is essentially plain denoising process that resorts to the diffusion models to generate final composited outcomes.

$\epsilon_\theta(\mathbf{x}_t, t, \varnothing)$ still produces significant reconstruction errors. Seeking for the cause of this phenomenon, [39] shows that any textual information described and associated within the input text prompt can induce a divergence between the backward and forward ODE trajectories. Therefore, [39] introduces a special exceptional prompt denoted as $\mathbf{P}_{exc}$ that completely eliminates all information by assigning a common value for all token numbers and discard positional embeddings for the input prompt as illustrated in Fig. 4. We directly employ this inversion technique to compute noisy latents of input non-tumorous background and tumour object images for precise reconstruction and composition intertwine.

4.3. Composition Generation

After inverting the non-tumorous background and tumour object images into their respective noisy latent codes $\mathbf{x}_T^{bg}$ and $\mathbf{x}_T^{obj}$, we propose a novel mechanism to initiate the composition process at a selective timestep which is simple yet effective in generating desired composited results.

Initial Timestep Selection. To construct starting point $\mathbf{x}_T^{comp}$ for initiating composition process, [39] merges the inverted latents $\mathbf{x}_T^{bg}$ and $\mathbf{x}_T^{obj}$ with another noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via below formulation:

$$\mathbf{x}_T^{comp} = \mathbf{x}_T^{obj} \odot \mathbf{M}_{obj} + \mathbf{x}_T^{bg} \odot (\mathbf{1} - \mathbf{M}_{user}) + \mathbf{z} \odot (\mathbf{M}_{obj} \oplus \mathbf{M}_{user}), \quad (7)$$

where $\mathbf{1} - \mathbf{M}_{user}$ indicates region outside user mask, and $\mathbf{M}_{obj} \oplus \mathbf{M}_{user}$ is the XOR between the object segmentation mask and user mask, denoting the transition area. Subsequently, they conduct composition process that concurrently involves denoising $\mathbf{x}_T^{comp}$, $\mathbf{x}_T^{bg}$, and $\mathbf{x}_T^{obj}$, in which they introduce a mechanism to extract the self-attention maps of object and background images from the diffusion models then inject them into those of composited results. Those this operation has demonstrated to be effective in injecting the object into the user-specified area, the object appearance tends to deviate from the original and loses its identifying features as in Fig. 6. Furthermore, randomness in initializing values within transition area may induce unwanted artifacts.

To deal with these issues, we propose a simple yet effective technique that allows for precise reconstruction and incorporation of tumour objects into background scans with preserved identity and seamless blending quality. Specifically, we delay the initiation of composition process until a selected timestep $0 < T' < T$ and slightly modify the formulation in Eq. 7 to construct the starting point as:

$$\mathbf{x}_{T'}^{comp} = \mathbf{x}_{T'}^{obj} \odot \mathbf{M}_{obj} + \mathbf{x}_{T'}^{bg} \odot (\mathbf{1} - \mathbf{M}_{obj}), \quad (8)$$

The intuition behind is that the more timesteps the denoising progress, the more information are reconstructed in $\mathbf{x}_{T'}^{bg}$ and $\mathbf{x}_{T'}^{obj}$, hence the more they can be incorporated to $\mathbf{x}_{T'}^{comp}$. Additionally, thanks to the rich prior and powerful denoising capability of pretrained text-to-image diffusion model,

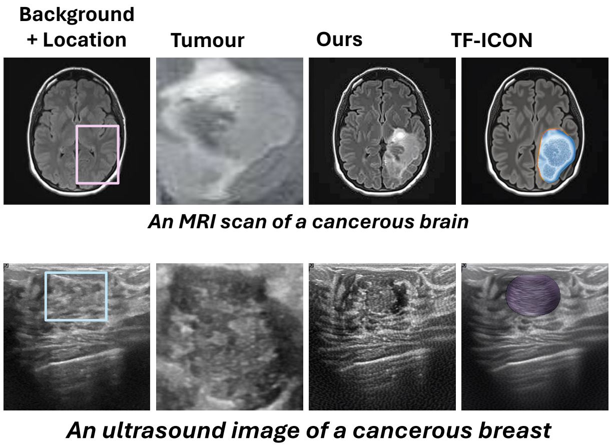
Background + Location
Tumour
Ours
TF-ICON
An MRI scan of a cancerous brain
An ultrasound image of a cancerous breast

Figure 6. TF-ICON suffers from object identity loss and noticeable artifacts while our approach does not.

$\mathbf{x}_{T'}^{comp}$ can be gradually refined to produce seamless composited results without necessitating $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for smooth transition areas and self-attention maps injection mechanism for effective composition as claimed in [39].

Background Preservation Trick. It is to note that the incorporation in Eq. 8 may affect the reconstruction of background area outside user mask. Therefore, to balance the generation of tumour object within user mask and preservation of background details, we introduce a threshold τ to regulate the trajectory recification process which entails replacing the background area outside user mask of composited image with the reconstructed background image at various timesteps similar to [39]. Formally, such process can be expressed as:

$$\mathbf{x}_t^{comp} = \mathbf{x}_t^{comp} \odot \mathbf{M}_{user} + \mathbf{x}_t^{bg} \odot (\mathbf{1} - \mathbf{M}_{user}), \quad (9)$$

where $t \in [\tau, T']$. Notably, only applying this process at final step can lead to unwanted artifacts as demonstrated in [39].

5. Experiments

5.1. Setup

Benchmark Dataset. Since there is currently no available dataset with the objective of evaluating medical cancer image composition, we construct a new benchmark dataset via a data acquisition process involving two public datasets named Br35H and BUSI acquired from Kaggle. Br35H is designated for Brain Tumor Detection 2020 challenge which comprises of 1500 non-tumorous and 1500 tumorous images with human-annotated segmentation mask of the tumour. Meanwhile, BUSI targets at breast cancer detection that provides more than 1500 breast ultrasound images. As our proposed approach follows training-free manner, it is sufficient to be assessed by a relatively small benchmark. Therefore, we collect 400 samples in total with 300 from Br35H and 100 from BUSI for evaluation. Each sample contains a tumour-free background image, a tumour object image with its segmentation mask, a user-specified mask of which location and size are generated in accordance to the background and object sizes, and a fixed input text prompt "An MRI scan of a cancerous brain" for samples originating from Br35H and "An ultrasound image of a cancerous breast" for those from BUSI.

Implementation Details. We employ the preprocessing process from TF-ICON [39] to rescale tumour object image and relocate it according to the user mask. Then, we conduct composition experiments using our proposed method as depicted in Fig. 5. First, we leverage exceptional prompt inversion to invert non-tumorous background and tumour object images into their respective latent codes $\mathbf{x}_T^{bg}$ and $\mathbf{x}_T^{obj}$ with $T = 20$. Then, we perform composition starting from $T' = 6$ and set $\tau = 3$ to retain background details. We conduct all experiments on NVIDIA Geforce RTX 3090 GPUs with a fixed seed for fair comparisons.

5.2. Comparison

We evaluate the performance of our method against prior state-of-the-art training-free frameworks including SDEdit [40], Blended Diffusion [3], and TF-ICON [39]. Note that we also include a naive composition approach involving simply copy and paste the tumour object onto the clean background image.

Qualitative Results. As shown in Fig. 7, our framework demonstrates the ability to seamlessly compose tumour objects into diverse background images while preserving their inherent identities and inducing no noticeable artifacts in transition area. In contrast, we can observe that the naive Copy-Paste approach always leads to sharp boundaries around the incorporated tumours. For SDEdit, the background of composited results are not well-preserved. Meanwhile, Blended Diffusion and TF-ICON fall short in retaining the shape and structure of the tumours, and their blending ability is inferior with critical color disparity in the tumours appearance.

Quantitative Results. We quantitatively assess the performance of our method against other approaches considering four metrics: (1) LPIPS$_{bg}$ [66] to assess background consistency, (2) LPIPS$_{fg}$ [66] to evaluate low-level similarity between the edited region and the tumour image, (3) CLIP$_{Image}$ [44] to assess semantic similarity between the edited region and the tumour in the CLIP embedding space, and (4) CLIP$_{Text}$ [44] to measure semantic alignment between the text prompt and the composited image. As demonstrated in Tab. 1, our method significantly outperforms other baselines on blending quality of the tumour in the composited image and preserving its identifying features which is consistent with the qualitative evaluation. It is also noted

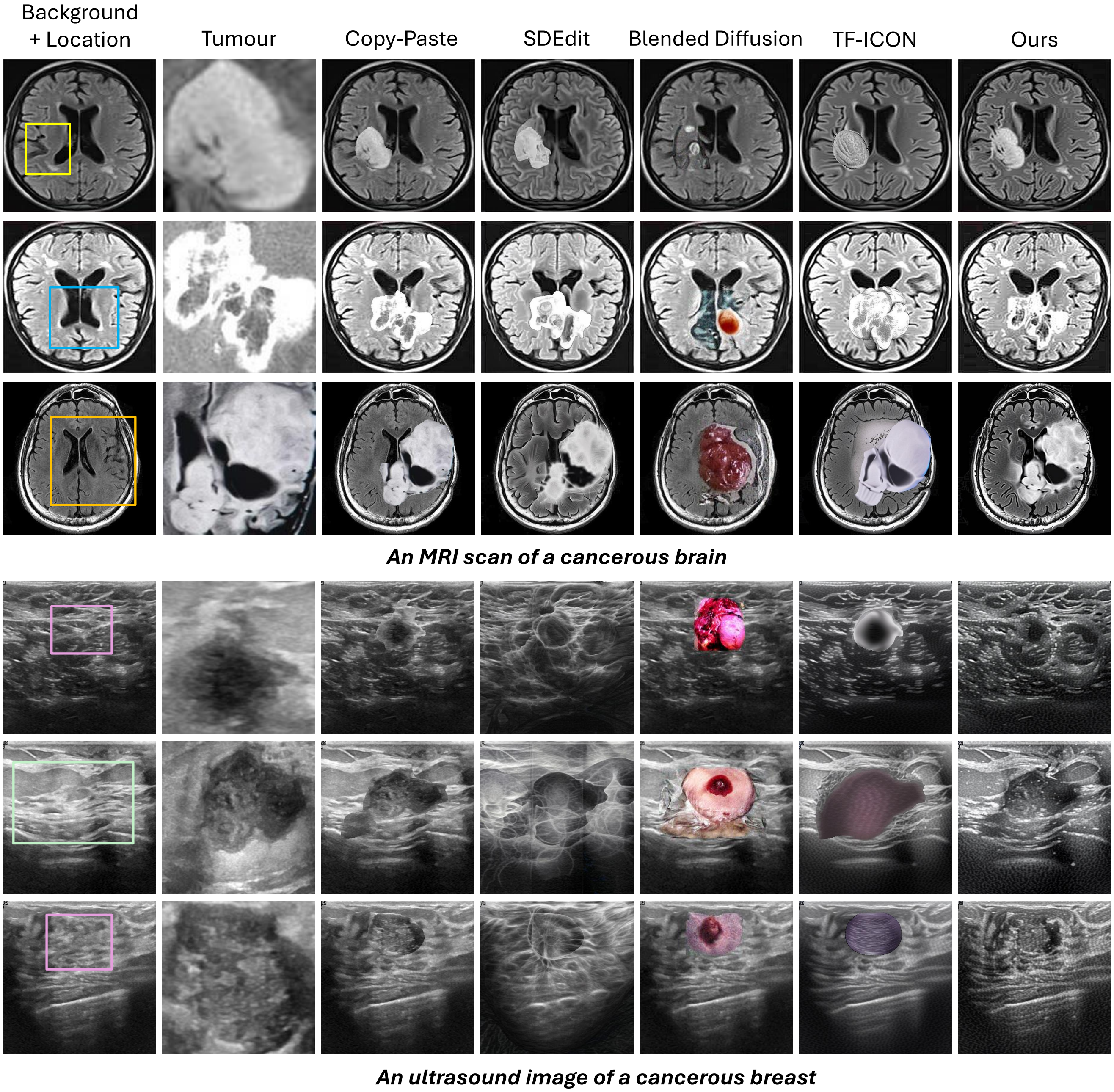
Background + Location
Tumour
Copy-Paste
SDEdit
Blended Diffusion
TF-ICON
Ours
An MRI scan of a cancerous brain
An ultrasound image of a cancerous breast

Figure 7. Qualitative comparison among our method and other prior SOTA training-free frameworks. The upper and lower groups respectively are composited results obtained from Br35H and BUSI samples.

that while Blended Diffusion can effectively retain the background information, it performs poorly in incorporating the tumour objects into composition. Meanwhile, SDEdit can generate composited results well-aligned with input text prompt similar to Blended Diffusion, yet it can barely maintain the background.

5.3. Ablation Studies

Selection of T'. We experiment with different values for T' and show qualitative results in Fig. 8 as well as quantitative results in Tab. 2 for comparison. We can observe that the lower the value for T', the more background and tumour object information are reconstructed and incorporated, eventually the better and more realistic the final compos-

Method	Br35H				BUSI			
	LPIPS$_{bg}$ ↓	LPIPS$_{fg}$ ↓	CLIP$_{Text}$ ↑	CLIP$_{Image}$ ↑	LPIPS$_{bg}$ ↓	LPIPS$_{fg}$ ↓	CLIP$_{Text}$ ↑	CLIP$_{Image}$ ↑
Blended [3]	0.0469	0.6584	30.9963	77.2117	0.138	0.7065	32.5691	81.7168
SDEdit [40]	0.2567	0.4502	31.7521	85.0899	0.4876	0.5746	31.1851	82.4449
TF-ICON [39]	0.1474	0.4859	31.0814	84.7188	0.2558	0.4655	30.9683	81.9469
Ours	**0.1738**	**0.4128**	**31.186**	**87.3847**	**0.2855**	**0.4027**	**30.7405**	**87.7517**

Table 1. Quantitative performance achieved by our method compared to prior training-free frameworks. Our results are shown in bold and the best results are in red.

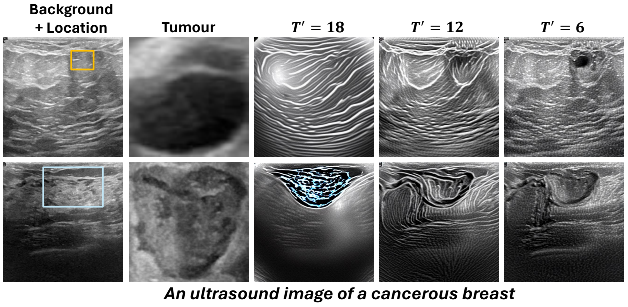

Figure 8. Ablation Study: Different selections of T'

Config	LPIPS$_{bg}$ ↓	LPIPS$_{fg}$ ↓	CLIP$_{Text}$ ↑	CLIP$_{Image}$ ↑
$T' = 18$	0.4384	0.5947	23.7741	80.4519
$T' = 12$	0.3743	0.5385	23.5959	80.7465
$T' = 6$	0.2855	0.4027	30.7405	87.7517

Table 2. Ablation Study: Different selections of T'

ited outcomes. These results are consistent with the intuition mentioned in **Initial Timestep Selection** Section 4.3.

Selection of τ. We fix $T' = 6$ and additionally experiment with different values for τ to assess its effects in retaining the background information. As demonstrated in Fig. 9, $\tau = 3$ attains the best balance among background details preservation, robustness against noises, and visual contrast. Meanwhile, $\tau = 1$ gives results with the best visual contrast but are the least to resemble the original background, and $\tau = 5$ produces the closest background but with lowest visual contrast.

6. Conclusion

In this work, we explore a new topic dubbed medical cancer image composition in which we present a novel training-free diffusion-based approach for high-quality and seamless image-guided compositional outcomes. Our method comprises of two stages respectively are exceptional prompt image inversion and composition generation with designated initial timestep selection that operates in synergy to guide composition process in producing desired results. Our experimental results demonstrate the potentiality and effectiveness of our method against several other training-free frameworks. We hope that this work can inspire future research on this topic.

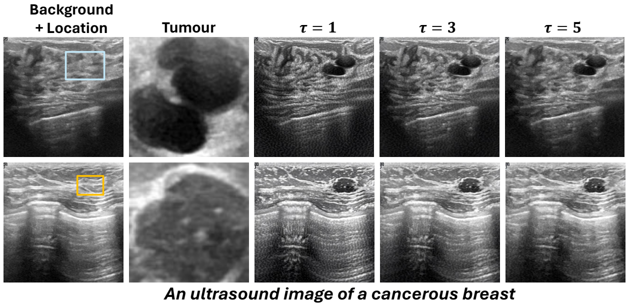

Figure 9. Ablation Study: Different values of τ

References

[1] Roa'a Al-Emaryeen, Sara Al-Nahhas, Fatima Himour, Waleed Mahafza, and Omar Al-Kadi. Deepfake image generation for improved brain tumor segmentation. In *2023 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)*. IEEE, 2023. 3

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, 2022. 3

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), 2023. 3, 6, 8

[4] Samaneh Azadi, Deepak Pathak, Sayna Ebrahimi, and Trevor Darrell. Compositional gan: Learning image-conditional binary composition, 2019. 3, 4

[5] Burak Beynek, Şebnem Bora, Vedat Evren, and Aybars Ugur. Synthetic skin cancer image data generation using generative adversarial neural network. *International Journal of Multidisciplinary Studies and Innovative Technologies*, 5(2):147–150, 2021. 3

[6] Junyan Cao, Yan Hong, and Li Niu. Painterly image harmonization in dual domains, 2023. 3

[7] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick

Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023. 2

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 3

[9] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8407–8416, 2019. 3

[10] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification, 2020. 3

[11] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance, 2022. 2, 4

[12] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020. 3

[13] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 4

[14] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022. 2

[15] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1310–1319, Los Alamitos, CA, USA, 2017. IEEE Computer Society. 3

[16] Yuheng Fan, Hanxi Liao, Shiqi Huang, Yimin Luo, Huazhu Fu, and Haikun Qi. A survey of emerging applications of diffusion probabilistic models in mri, 2023. 3

[17] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis, 2023. 3

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 2

[19] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based domain tuning for fast personalization of text-to-image models, 2023. 2

[20] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3

[21] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022. 2, 4

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 4

[23] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Shadow generation for composite image in real-world scenes, 2022. 3

[24] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021. 3

[25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. 4

[26] Ranpreet Kaur, Hamid GholamHosseini, and Roopak Sinha. Synthetic images generation using conditional generative adversarial network for skin cancer classification. In *TENCON 2021 - 2021 IEEE Region 10 Conference (TENCON)*, pages 381–386, 2021. 3

[27] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Conference on Computer Vision and Pattern Recognition 2023*, 2023. 2

[28] Benjamin L. Kidder. Advanced image generation for cancer using diffusion models. *bioRxiv*, 2023. 3

[29] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation, 2022. 2, 4

[30] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023. 2

[31] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023. 2, 4

[32] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing, 2018. 3

[33] Daquan Liu, Chengjiang Long, Hongpan Zhang, Hanning Yu, Xinzhi Dong, and Chunxia Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8136–8145, 2020. 3

[34] Liu Liu, Zhenchen Liu, Bo Zhang, Jiangtong Li, Li Niu, Qingyang Liu, and Liqing Zhang. Opa: Object placement assessment dataset. *arXiv preprint arXiv:2107.01889*, 2021. 3

[35] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds, 2022. 4

[36] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models, 2023. 3

[37] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 4

[38] Lingxiao Lu, Jiangtong Li, Junyan Cao, Li Niu, and Liqing Zhang. Painterly image harmonization using diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 2023. 3

[39] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. Tf-icon: Diffusion-based training-free cross-domain image composi-

tion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2294–2305, 2023. 2, 4, 5, 6, 8

[40] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations, 2022. 2, 6, 8

[41] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition, 2024. 3

[42] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation, 2023. 2, 4

[43] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery, 2021. 2

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 6

[45] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 2, 3

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 4

[47] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2

[49] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. 4

[50] Yichen Sheng, Jianming Zhang, and Bedrich Benes. Ssn: Soft shadow network for image compositing, 2021. 3

[51] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. 3

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 2, 4

[53] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022. 4

[54] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18310–18319, 2023. 2

[55] R. Toda, A. Teramoto, M. Kondo, K. Imaizumi, K. Saito, and H. Fujita. Lung cancer CT image generation from a free-form sketch using style-based pix2pix for data augmentation. *Sci Rep*, 12(1):12867, 2022. 3

[56] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Trans. Graph.*, 40(4), 2021. 4

[57] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, and Visesh Chari. Learning to generate synthetic data via compositing, 2019. 3

[58] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation, 2022. 2, 4

[59] Sam Witteveen and Martin Andrews. Investigating prompt engineering in diffusion models, 2022. 3

[60] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending, 2019. 3

[61] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization, 2022. 3

[62] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. *arXiv preprint arXiv:2211.13227*, 2022. 2

[63] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation, 2022. 2

[64] Lingzhi Zhang, Tarmily Wen, and Jianbo Shi. Deep image blending, 2019. 3

[65] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII*, page 566–581, Berlin, Heidelberg, 2020. Springer-Verlag. 3

[66] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[67] Shuyang Zhang, Runze Liang, and Miao Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105 – 115, 2019. 3