electronics

MDPI

Article

Low-Cost Training of Image-to-Image Diffusion Models with Incremental Learning and Task/Domain Adaptation

Hector Antona, Beatriz Otero and Ruben Tous *

Department of Computer Architecture, Universitat Politècnica de Catalunya, Jordi Girona 1-3, 08034 Barcelona, Spain; hector.antona@estudiantat.upc.edu (H.A.); beatriz.otero@upc.edu (B.O.)
* Correspondence: ruben.tous@upc.edu; Tel.: +34-93-405-4044

Abstract: Diffusion models specialized in image-to-image translation tasks, like inpainting and colorization, have outperformed the state of the art, yet their computational requirements are exceptionally demanding. This study analyzes different strategies to train image-to-image diffusion models in a low-resource setting. The studied strategies include incremental learning and task/domain transfer learning. First, a base model for human face inpainting is trained from scratch with an incremental learning strategy. The resulting model achieves an FID score almost equivalent to that of its batch learning equivalent while significantly reducing the training time. Second, the base model is fine-tuned to perform a different task, image colorization, and, in a different domain, landscape images. The resulting colorization models showcase exceptional performances with a minimal number of training epochs. We examine the impact of different configurations and provide insights into the ability of image-to-image diffusion models for transfer learning across tasks and domains.

Keywords: diffusion probabilistic models; deep learning; adaptive learning; transfer learning; image inpainting; image colorization; image-to-image translation; training efficiency

Citation: Antona, H.; Otero, B.; Tous, R. Low-Cost Training of Image-to-Image Diffusion Models with Incremental Learning and Task/Domain Adaptation. *Electronics* **2024**, *13*, 722. https://doi.org/10.3390/electronics13040722

Received: 29 December 2023
Revised: 2 February 2024
Accepted: 7 February 2024
Published: 10 February 2024

1. Introduction

The present work addresses some of the challenges associated with diffusion models specialized in image-to-image translation tasks, such as inpainting and colorization. We give particular emphasis to the findings of "Palette: Image-to-Image Diffusion Models" [1], a widely recognized method with a multi-task generalist approach. Palette introduces a unified multi-task framework for image-to-image translation. Palette's approach leverages the versatility of diffusion models (DMs) and demonstrates superior performance compared to strong generative adversarial networks (GANs) and regression baselines. However, while multi-task generalist approaches offer versatility, it is essential to consider their limitations. In certain scenarios, specialized models tailored specifically to individual tasks may achieve superior performance, and that is what we aim to reproduce. Additionally, the training process of a generalist model based on DMs requires substantial computational resources and extensive datasets, posing challenges in terms of feasibility and accessibility for developers [2].

In light of these considerations, this study aims to provide a practical approach that mitigates the challenges associated with image-to-image diffusion models in low-resource settings, with a special emphasis on multi-task generalist approaches. Specifically, the focus is on reducing the computational requirements and training time by applying incremental learning and task/domain transfer learning techniques. We utilize a concrete task, such as human face inpainting, as the starting point. A base model for this task is trained from scratch with an incremental learning strategy. We compare the performance achieved by this model with its batch-learning equivalent. Second, the base model is fine-tuned to perform a different task, image colorization. We compare the computational cost and the performance of the resulting model with a multi-task approach and with a specialized

method. Finally, the study explores an alternative approach that aims to adapt a pre-trained model from one domain (faces) to another (landscapes). We assess the extent to which the knowledge embedded in the pre-trained face colorization model can be utilized to efficiently colorize landscape images with minimal training. We examine the impact of different configurations and provide insights into the application of a domain transfer learning approach to this type of model.

Overall, this work contributes to the advancement of DMs by addressing their high computational requirements and training time [3]. By providing solutions for training efficiency and knowledge transfer, the proposed approaches enhance the feasibility and accessibility of DMs for various image-to-image translation tasks and data domains.

2. Diffusion Models Overview

DM architecture consists of two Markov chains [2]. The first one incorporates noise in its input data at each timestep, making it converge toward a simpler data distribution (generally Gaussian noise). This one is carefully built to inject the noise so that the data truly converge toward real Gaussianity. Because of this, no parameters are usually trained in this half. The second Markov chain is in charge of reversing the noise injection the previous chain incorporated. This process is achieved by training many transition models, one for each timestep, where each gradually denoises the data. Once this architecture is built and trained, the generation of new samples is obtained through the input of real Gaussian noise (not obtained from the first chain); then, it is forwarded through the second chain and obtains the newly generated data sample.

Figure 1 shows how the input image $x(0)$ is forwarded through all timesteps until reaching the final state with a Gaussian noise distribution in $x(T)$. Then, it repeats all the steps backward until it reconstructs the original image.

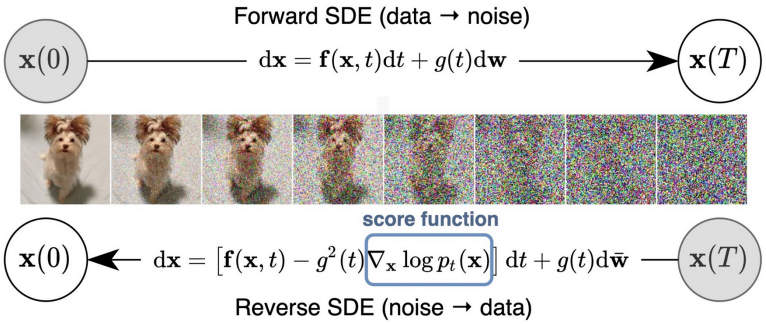

Figure 1. Diffusion model architecture from [4], transforming data into a simple noise distribution. Then, this process can be reversed by utilizing the score of the distribution at each intermediate timestep.

In terms of the model's architecture, it is important to note that Ho et al. [5] utilized a U-Net, ensuring that the input and output of the model were of the same size. In essence, a U-Net is a symmetric architecture that incorporates skip connections between encoder and decoder blocks, enabling the preservation of feature information [6]. A schematic of its architecture is shown in Figure 2. Typically, the input image undergoes downsampling and subsequent upsampling operations until it reaches its original size. In the original implementation of denoising diffusion probabilistic models, the U-Net architecture consists of wide ResNet blocks, group normalization, and self-attention blocks [5]. Additionally, to specify the diffusion timestep, a sinusoidal position embedding is incorporated into each residual block.

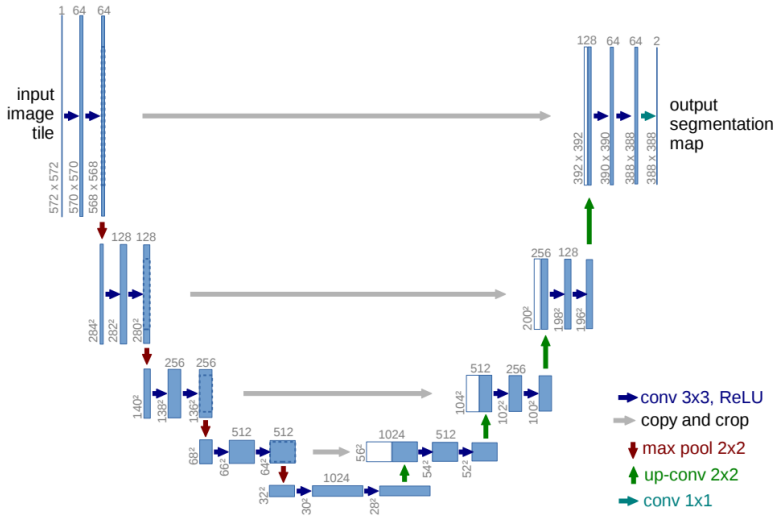

Figure 2. U-Net architecture (extracted from [6]). Each blue box represents a feature map with multiple channels, and the number of channels is indicated on top of the box. The x–y size of the feature map is provided at the lower left edge of the box. White boxes indicate copied feature maps, while the arrows indicate the various operations performed.

2.1. Three Directions of Improvement

However, we can identify three main issues and areas of improvement for DMs: sampling-acceleration enhancement, likelihood-maximization enhancement, and data-generalization enhancement.

2.1.1. Sampling Efficiency

The first issue that DMs present is that, to achieve the desired results, the Markov chains that build them must have a very large number of steps, with values around 1000 steps, in order to ensure one obtains high-quality samples. Each of these steps requires evaluating a neural network model once. Adding up all steps, forwarding every corresponding neural network demands long training times and loads of computational operations. All this results in obtaining substantially slower models than GANs, which are also analyzed later, for example, in [3]. Because of this high number of small networks that DMs require for training, both training and evaluation processes take a very long time, needing a lot of computational power too. Later, this work specifically tackles this issue as its main focus, as previously mentioned.

2.1.2. Likelihood Maximization

Given how DMs work, their main objective is to maximize the log-likelihoods of the data they generate, but log-likelihood is not directly optimized by the weighted combination of score-matching losses. This is equivalent to minimizing the divergence between the forward and reverse processes through the whole Markov chain. This is defined as a variational lower bound (VLB). However, this VLB easily finds sub-optimal log-likelihoods. Different techniques exist to solve the issue, such as noise schedule optimization, reverse variance learning, and exact log-likelihood evaluation [2].

Noise schedule optimization: When defining and building diffusion models, noise injection in the forward process is programmed without trainable parameters. By optimizing the forward noise schedule of diffusion models, one can further maximize the VLB in order to achieve higher log-likelihood values [7].

Reverse variance learning: Classical definitions of DMs assume that transitions between timesteps in the reverse Markov chain have pre-established variance parameters. In addition to maximizing the VLB, some techniques propose training the reverse variances. Improved denoising diffusion probabilistic models (iDDPMs) [8] propose learning the

reverse variances by parameterizing them with linear interpolation and training them using a hybrid objective. This results in higher log-likelihoods and faster sampling without losing sample quality.

Exact log-likelihood evaluation: This last technique is straightforward but more computation-heavy, as it aims to extract the log-likelihood formulation by solving the reverse stochastical differential equation of the model [4].

2.1.3. Data Generalization

DMs have achieved great success in fields such as computer vision. However, within some other subjects, they do not seem to work equally well. Many important data domains have special structures that must be taken into account to achieve correct DM function. As an example, models that heavily rely on score functions defined on continuous data domains pose a problem for this kind of model. The text-to-image synthesis also exemplifies this situation. To solve this scenario, some adjustments to the basic architecture of the diffusion model have to be made to fit and adapt to the discrete data space [9]. On the other hand, when working on supported data domains, such as image-to-image translation, where they operate with no need for any data adaption, DMs showcase excellent results and superior capabilities over other generative architectures we analyze in the following section.

2.2. State of the Art

Diffusion probabilistic models are outperforming existing state-of-the-art methods in various fields and applications [10,11]. Recent advancements in image-to-image translation tasks, such as inpainting and colorization, have showcased the potential of diffusion models to achieve superior performance compared to other state-of-the-art generative architectures. However, they are mostly focused on one target domain and single-task solving.

The RePaint [12] paper introduced a novel approach to free-form inpainting, which involves filling in missing regions in an image based on an arbitrary binary mask. Unlike existing methods that are limited to specific mask distributions and often produce texturally simplistic results, RePaint utilizes a pre-trained unconditional denoising DM as a generative prior. Instead of training a mask-conditional model, RePaint conditions the generation process by sampling from the known regions of the image during reverse diffusion iterations. This allows the model to produce high-quality and diverse output images for inpainting tasks with any mask type. The paper presented experiments on the face and general-purpose image inpainting, demonstrating that RePaint outperforms state-of-the-art autoregressive and GAN-based approaches (e.g., [13,14]) for a variety of mask distributions. The proposed method leverages the power of DMs for semantically meaningful generation and texture synthesis, and it provides an effective conditioning strategy for inpainting. While using a problem-specific algorithm can help reduce training costs, it often comes at the expense of versatility. These algorithms are specifically designed to solve a particular problem and may struggle to adapt to new or different tasks. Their rigid structure limits their ability to generalize and apply knowledge to other domains, which is what we are aiming to solve in a more generalized and versatile way. On the other hand, COPAINT [15] also applies a denoising DM to the inpainting problem but addresses the incoherence between revealed and unrevealed regions with a Bayesian framework, approximating the posterior distribution in a way that allows the errors to gradually drop to zero throughout the denoising steps, thus strongly penalizing any mismatches with the reference image.

Concerning the state of the art in DMs related to colorization, we can find the Palette [1] paper. It presented a unified framework for image-to-image translation using conditional diffusion models. The paper focused on four challenging tasks: colorization, inpainting, uncropping, and JPEG restoration. The authors demonstrated that their implementation of image-to-image diffusion models surpasses strong GAN and regression baselines in all tasks without requiring task-specific customization or advanced techniques. They explored

the impact of different loss functions and neural network architectures, emphasizing the importance of self-attention layers. The paper also introduced a standardized evaluation protocol based on ImageNet, incorporating human evaluation and various sample quality scores. The authors advocated for the versatility and generality of diffusion models in image manipulation and highlighted the performance of a generalist, multi-task diffusion model compared to task-specific specialist counterparts. The paper concluded with an evaluation of Palette on the colorization task, where it achieved state-of-the-art results and demonstrated close resemblance to the original images in terms of quantitative metrics and human evaluation.

It is important to highlight that Palette [1] employs a multi-task model that offers increased versatility but does not specialize in any specific task but in many. Although our foundation lies in Palette, a multi-task model, our proposed method adopts a single-task transfer learning approach. In contrast to the parallel learning of shared information across multiple tasks, our objective is to sequentially utilize knowledge from a source task to enhance the training of a target task. By training and adapting the model from a baseline one, it can bend and optimize its performance based on the specific requirements of the target domain and task. This flexibility enables researchers to tailor the model to different applications and explore a wide range of image-to-image translation tasks while still achieving competitive performance and having low training time requirements.

Other works have also tried to compensate for the limitations of DMs. The paper "Wavelet Diffusion Models are Fast and Scalable Image Generators" [16] addresses the challenge of slow training and inference speeds in diffusion models despite their high-quality image generation capabilities. The authors proposed a novel wavelet-based diffusion scheme that leverages wavelet decomposition to extract low- and high-frequency components from images and features. This approach significantly reduced processing time while maintaining good generation quality. Experimental results on various datasets demonstrated that the proposed wavelet diffusion framework bridges the speed gap between diffusion models and GANs, making diffusion models more suitable for real-time and large-scale applications.

3. Methods

This work is inspired by "Palette: Image-to-Image Diffusion Models" [1]. This paper presents a novel unified framework for image-to-image translation tasks, such as colorization, inpainting, uncropping, and JPEG restoration. Palette's proposed approach leverages conditional diffusion models and achieves superior performance compared to strong GAN and regression baselines without requiring task-specific hyper-parameter tuning, architectural customization, or additional sophisticated techniques. The authors introduced a straightforward implementation of image-to-image diffusion models that outperforms existing methods across all considered tasks. Overall, the presented unified framework for image-to-image translation utilizing conditional diffusion models demonstrates remarkable performance, emphasizes the importance of self-attention mechanisms, introduces a standardized evaluation protocol, and showcases the effectiveness of a generalist approach in multi-task scenarios. Palette is an image-to-image diffusion model, i.e., a conditional diffusion model of the form $p(\boldsymbol{y} \mid \boldsymbol{x})$, where both $\boldsymbol{x}$ and $\boldsymbol{y}$ are images (e.g., $\boldsymbol{x}$ is a grayscale image, and $\boldsymbol{y}$ is a color image). A conditional diffusion model converts samples from a standard Gaussian distribution into samples from an empirical data distribution through an iterative denoising process conditional on an input signal. The denoising loss function is obtained by training a neural network f_θ. Given a training output image $\boldsymbol{y}$, a noisy version $\widetilde{\boldsymbol{y}}$ is generated:

$$\widetilde{\boldsymbol{y}} = \sqrt{\gamma}\,\boldsymbol{y}_0 + \sqrt{1-\gamma}\,\boldsymbol{\epsilon}\,, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{1}$$

f_θ is trained to predict the noise vector $\boldsymbol{\epsilon}$, i.e., to denoise $\widetilde{\boldsymbol{y}}$ given $\boldsymbol{x}$ and a noise level indicator γ, for which the loss is

$$\mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})}\mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,I)}\mathbb{E}_{\gamma}\left\|f_{\theta}(\boldsymbol{x},\underbrace{\sqrt{\gamma}\,\boldsymbol{y}+\sqrt{1-\gamma}\,\boldsymbol{\epsilon}}_{\tilde{\boldsymbol{y}}},\gamma)-\boldsymbol{\epsilon}\right\|_{p}^{p} \tag{2}$$

Inference is performed via the learned reverse process. Since the forward process is constructed so the prior distribution $p(y_T)$ approximates a standard normal distribution $\mathcal{N}(\boldsymbol{y}_T|\mathbf{0},\boldsymbol{I})$, the sampling process can start at pure Gaussian noise, followed by T steps of iterative refinement.

Figure 3 graphically shows the workflow of the inference process, which is a reverse Markovian process that starts with a pure Gaussian noise image y_t and a condition image x (the target image with a missing region in the inpainting task) and iteratively performs a conditional denoising with the U-Net model f_θ.

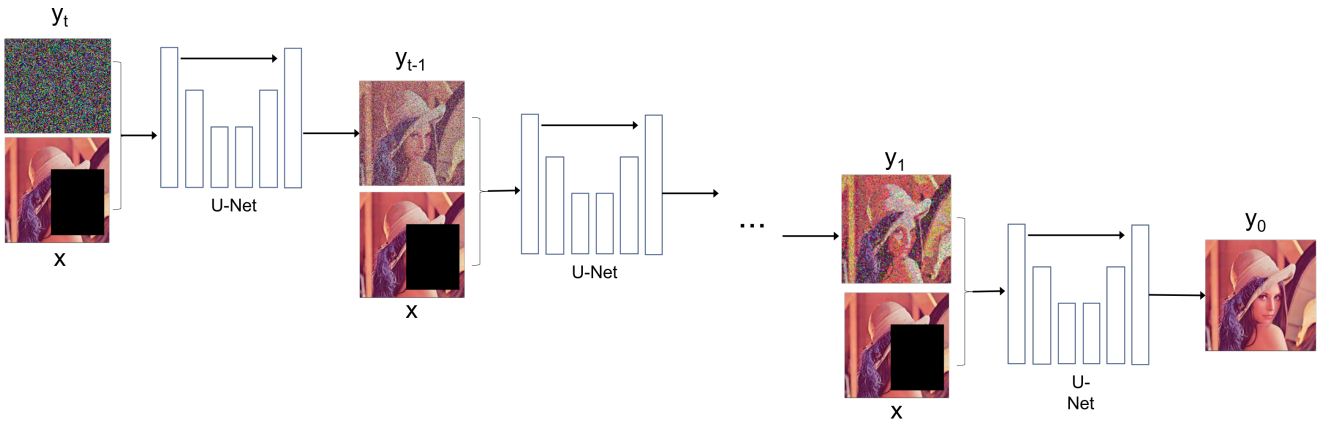

Figure 3. Conditional (image-to-image) diffusion model inference workflow.

Palette's architecture employs U-Net, initially introduced by Ho et al. [5], but with significant modifications. The network's structure is based on the 256 × 256 class-conditional U-Net model proposed by Dhariwal and Nichol [17]. However, two major differences set their architecture apart: first, they do not use class conditioning, and second, Palette incorporates additional conditioning of the source image through concatenation, following the approach of Saharia et al. [18].

In this study, we focus on the drawbacks associated with utilizing a multi-task generalist approach for image-to-image translation. While this approach offers versatility, it is important to consider its potential limitations. In certain scenarios, employing a specialized model tailored specifically to a given task may yield superior performance compared to a generalist model. Moreover, the training process of the generalist model necessitates significant computational resources and extensive datasets, thus presenting challenges in terms of feasibility and accessibility for potential developers [19]. Through the examination of the advantages and limitations of employing a multi-task generalist approach for image-to-image translation, this paper provides a practical approach that requires an easier and computationally lighter training process. Figure 4 provides a graphical overview of the methodology. First, a base model for human face inpainting is trained from scratch with an incremental learning strategy. Second, the base model is fine-tuned to perform a different task, image colorization. The resulting colorization model is, in turn, fine-tuned to operate in a different domain, landscape images.

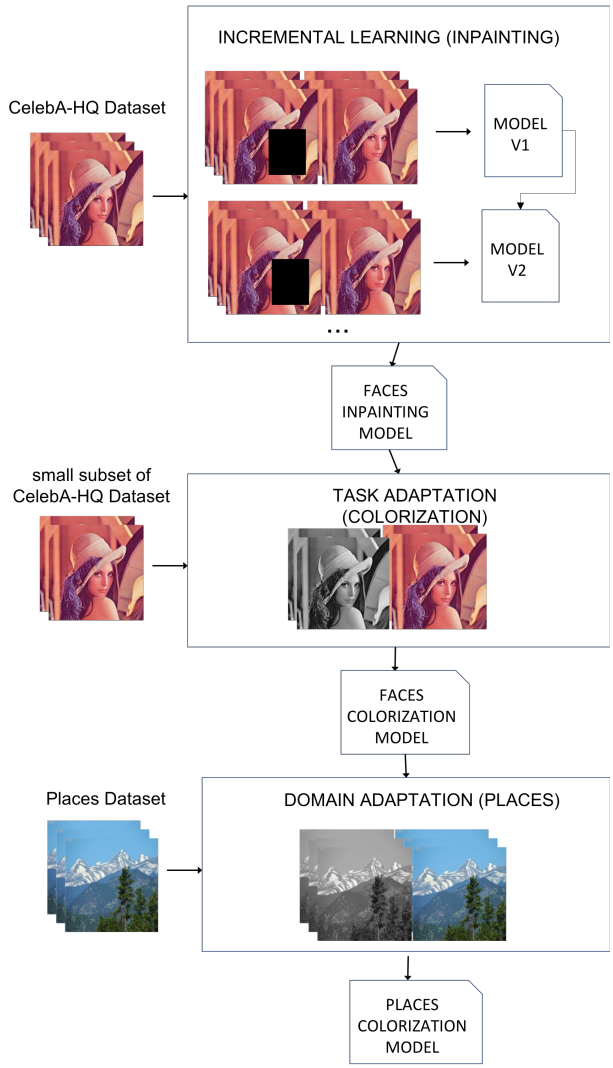

Figure 4. Outline of the methodology workflow.

3.1. Baseline—Inpainting

The study begins with training an image-to-image diffusion model for human face inpainting based on the Palette architecture. This is the base model that is later adapted for other purposes. Because of its complex requirements, image inpainting has been selected as the foundational task for the base model. This choice aims to leverage the acquired skills to subsequently achieve effectiveness in other tasks and domains with minimal cost. The study focuses on leveraging the CelebA-HQ dataset [20] to train and evaluate the proposed model.

The CelebA-HQ dataset is a high-resolution dataset, specifically focusing on human faces. It contains a large collection of high-quality images featuring various individuals, including celebrities and non-celebrities. The dataset covers a wide range of facial attributes, expressions, poses, and backgrounds, providing a diverse representation of human faces. Each image is carefully labeled with rich attribute annotations, enabling research in areas such as facial recognition, attribute prediction, and facial analysis. The CelebA-HQ dataset serves as a valuable resource for training and evaluating deep learning models that deal with facial image processing and analysis tasks like ours.

The base model aims to generate visually plausible and contextually coherent missing regions within the images. The utilization of the CelebA-HQ dataset during the training for this task enables the model to learn from diverse examples of human faces, thereby facilitating the acquisition of facial attributes, color patterns, and spatial relationships, very useful features that can be exploited in other tasks and applications. The base DM

architecture was obtained from an existing unofficial implementation of Palette [21], which was later modified and tuned to fit this work's scope. The existing code held all four Palette modules (colorization, inpainting, uncropping, and JPEG restoration), although only inpainting and uncropping were functional. Moreover, this first step's goal consisted solely of image inpainting. There were some pre-trained inpainting models that were used for the early steps of the project, but in the end, a long-trained model focused on inpainting was trained from scratch. The software in charge of this task was already built [21], with small changes required, but hyperparameter optimization was required to obtain better results, especially adjusting to training time restrictions. The code required a long training time (as expected) and required GPUs to run it. Because of this, we had to handle the parameters of the model to minimize the running time but keep a healthy balance so good results were still obtained. Initially, the model consisted of 2000 timesteps, which means that it added noise 2000 times and was denoised another 2000 times. We reduced the number of steps to 1200, as we experimentally found that quality results were still achieved while minimizing the timesteps. Also, to compensate for this modification, we proportionally increased the noise addition to end up with the same amount of distortion at the end of the Markov chain, passing from 0.9999 of maintained information between timesteps to 0.9983. It is worth mentioning that the limitation to the number of images per epoch fed to the model was one of the biggest factors that helped reduce the runtime, but this hyperparameter adjustment provided an extra speed-up.

Incremental Learning

We train all models with a minibatch size of 1024, a standard Adam optimizer, a 10k linear learning rate warmup schedule, and 0.9999 EMA. Algorithm 1 shows the training pseudocode inspired by [1] and described in Section 3.

Algorithm 1 Training a denoising model f_θ

1: **repeat**
2: $(\boldsymbol{x}, \boldsymbol{y}_0) \sim p(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x}$ is a grayscale image and $\boldsymbol{y}_0$ is a color image
3: $\gamma \sim p(y_T)$ (noise level indicator by T steps of iterative refinement)
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{y}_T|\boldsymbol{0}, \boldsymbol{I})$ (approximating a standard normal distribution)
5: Take a gradient descent step on
$\nabla_\theta \left\| f_\theta(\boldsymbol{x}, \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\epsilon}, \gamma) - \boldsymbol{\epsilon} \right\|_p^p$
6: **until** converged

As a diffusion model with more than 500M parameters, Palette is very expensive to train on the CelebA-HQ dataset. In order to enable training in a configuration with few resources, we adopted an incremental learning strategy. As a low-resource setting, we selected the free Google Colab tier. In contraposition to batch learning, incremental learning does not use the same set of images for all training epochs. We trained the model with episodic training sessions spanning 5 epochs each, which approximately aligns with the time limit imposed by Colab. After each session, the model's state was saved to ensure continuity. Subsequently, for the following 5 epochs, we reloaded the previously saved model state and continued training from that point onwards. This approach enabled us to train the model effectively while accommodating the time limitations imposed by the Colab platform. Yet, completing 5 epochs each session was not enough, as a total of 200 epochs were needed to obtain the desired results. For this reason, the learning rate and the number of timesteps were adjusted to reduce the running time.

In order to effectively monitor and reduce the duration of training epochs, it became necessary to address the number of utilized images. Employing a large number of images resulted in a significant exponential increase in training time, while excessively reducing the dataset posed the risk of detrimental effects, such as overfitting or subpar performance. To address this problem, a new training methodology was designed. The approach consisted of maintaining the whole dataset, but instead of accessing it all during each epoch, only a subset of it was considered. For the first epoch, a sub-selection of the images is considered.

After a few epochs, 3 to be precise, this subset was changed, having access to a different selection of data. Algorithm 2 shows the pseudocode of this incremental learning strategy.

Algorithm 2 Incremental training model

Require: N ←number of data subsets; E ←epochs per subset; D ←dataset; M ←model (new or pretrained)
1: Divide the training dataset D into N equal subsets $D^0, D^1, ..., D^{N-1}$
2: **for** $i \leftarrow 0$ to $N-1$ **do**
3: **for** $j \leftarrow 0$ to $E-1$ **do**
4: $M = train\_with\_Algorithm\_1(D^i, M)$ (three epochs)
5: **end for**
6: **end for**

The incremental learning model M was trained with new incoming data D^i applied to the M model obtained in the previous training phase. This approach facilitated shorter training times while maintaining access to a diverse range of data for learning purposes. Moreover, this strategy allowed each continuous stream to consume less memory and train models with limited resources.

To mitigate potential overfitting issues due to having fewer images for short periods of time, the model's dropout rate was moderately increased. These adaptations collectively enabled the training of 15 epochs per session, as opposed to the initial 5, resulting in notable improvements in overall training time without compromising the quality of the achieved outcomes.

The existing code already had a pre-trained model trained with 200 epochs as we did, although the pre-trained model would need a much longer training process than ours did. The results section further analyzes and compares our model with the pre-trained one, but with this faster and optimized approach, we obtained very similar results with a much shorter training process.

Reaching this point, we obtained a trustworthy base model, focused on the inpainting of human faces, that served as a departure point for all the following experiments.

3.2. Task Adaptation—Colorization

Now, this study investigates the adaptability of the long-trained inpainting model to the task of colorization. Colorization and inpainting are two challenging image-to-image translation tasks that entail different difficulties. Inpainting faces the problem of generating a semantically consistent structure, minimizing the incoherence between revealed and unrevealed regions. Colorization faces the problem of respecting the different semantic categories and producing high-fidelity colors. However, while distinct in their objectives, the basic premise suggests that the complex demands of inpainting likely involve knowledge that can be beneficially applied to colorization. The study delves into the learning dynamics of the model, shedding light on the process of acquiring colorization skills within the existing framework. As mentioned before, when learning from the CelebA-HQ dataset, the model acquired information about facial attributes, color patterns, and spatial relationships.

During the present stage, the model was provided with input data consisting of the original images altered by an incorporated square mask centered in a random point close to the center of the image, effectively occluding an area encompassing approximately one-fourth of the total image dimensions. Moreover, the mask consisted of random Gaussian noise. This approach ensured that the model would encounter a consistent and controlled input configuration, allowing it to focus specifically on the task of handling the masked region and generating accurate inpainting results. An example of the ground truth and the masked image is shown in Figure 5. This was modified to fit the new colorization task. The mask was no longer applied, and the image was converted to a gray-level scale. A relevant detail to highlight here is that DMs are known for keeping the same dimensionality of the data throughout the whole process, working with samples of size 256 × 256 and

3 channels as we worked with colored images. When obtaining the gray-scale image, we were cutting down the number of channels from three to one, which did not suit the architecture anymore. To solve this, the input to the model was adapted, obtaining the 1-channel grayscale image and passing it to a 3-channel image by having all three channels be the same. In Figure 6, we see the new input and ground truth images given to the model to train the new task.

With these modifications applied, we loaded the previously trained baseline model for inpainting and began training with the new grayscale-level input images to achieve colorization. The same technique of changing to different subsets of images every few epochs, as described in the previous section, was used again to maintain the speed of the training process. With only 20 epochs, the model already reached excellent results, showcasing an extremely short training period compared to the one needed for the inpainting task and demonstrating an improvement in efficiency when adapting the model to other tasks within the same data domain.

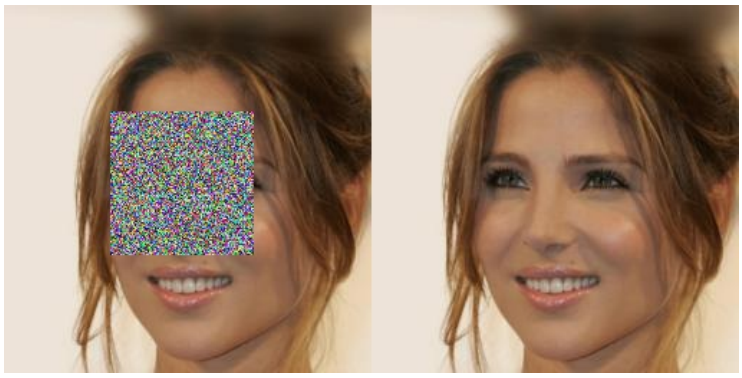

Figure 5. Masked image for inpainting and ground truth image.

Figure 6. Gray level (3 channels) and ground truth image.

3.3. Domain Adaptation—Landscapes Colorization

This study explores the DM's domain adaptation capabilities in addition to task adaptation. Up until now, we modified the task our model aimed to solve, passing from inpainting to colorization. This was a natural transition, as the DM learned the data features from the CelebA-HQ dataset, thus requiring a small training process to change from one task to another within the same scope. In this section, we explore the DM's capability of learning features inherent to the task instead of only learning features from the data. The idea behind this is to show that when you have a colorization model on a target data domain, the DM has learned how the fed data behaves and how to colorize it, but those two learned aspects are partially separable. That means that, when training the model to colorize the CelebA-HQ images, some part of the learning was dedicated exclusively to learning to colorize, independent of the data. Thus, being able to adapt to a second domain much faster is thanks to the acquired task-related knowledge.

In the previous section, we explore the opposite scenario. We assumed that the data structures were learned, and changing the task only required an extra effort. This time, we approach this from a less obvious perspective. To prove this, instead of changing the task of the model, we kept it and changed the data domain instead. We loaded the colorization model obtained in the previous section and retrained it with images from the Places-2 dataset, aiming to colorize them.

The Places-2 dataset [22] is a large-scale collection of images that encompasses a wide range of scenes and environments. It consists of diverse images captured from various locations worldwide, covering urban, rural, natural, and indoor settings. The dataset contains images depicting landscapes, buildings, interiors, streets, landmarks, and other visual elements commonly encountered in different geographic locations. It aims to provide a comprehensive representation of the visual characteristics and diversity of places, enabling research and development in areas such as scene understanding, object recognition, and image classification. In order to keep it simpler and require a smaller set of data overall, reducing its variance, we only considered images from the forest and mountain sets of the database.

Loading the previously trained model for face colorization, we retrained the model, this time feeding the Places-2 dataset's images as input but converted to a gray-level scale. Figures 7 and 8 show examples of the data fed for this experiment. Places-2 [22] has a high number of subsets of images featuring different kinds of places. In this experiment, only images from the folders 'Mountain' and 'Forest' were considered. We chose these images because they held some similarities, but we still had data from more than one class, increasing the generalization capabilities of our model. For the training, the images were treated the same way as in previous experiments. With only 25–30 epochs, the model already reached excellent results, showcasing an extremely short training period compared to the one needed when training a model from scratch.

Figure 7. Gray level and ground truth mountain image.

Figure 8. Gray level and ground truth forest image.

Making this differentiation between task and domain, we found that we could maintain one and change the other and, with the new setup, adapt the previous model to the newly defined one. In our case, we changed the task, exchanging inpainting with colorization, or we changed the domain by exchanging human faces with landscapes. However, these alterations have an endless number of possibilities. We could consider uncropping or image restoration as new tasks, and we could select datasets featuring animals or objects representing different domains, always making these changes one at a time.

4. Results

This section comprises several components. Firstly, we elaborate on the selection of our evaluation metric, the reasons why we chose it, and how it works. Subsequently, we present an analysis of each of the three distinct experiments conducted. The initial experiment involved the baseline inpainting model for human faces, followed by the task-adapted model for human face colorization. Finally, we examine the domain-adapted model specifically designed for landscape colorization. The performance and speed-up time were both analyzed for each experiment, providing a comprehensive assessment of their respective outcomes. Some examples of generated images are portrayed for qualitative results analysis. To quantitatively evaluate the performance of our image generation models, we adopted the Fréchet inception distance (FID) as a fundamental metric. FID has emerged as a widely accepted measure for evaluating the visual fidelity and diversity of generated images. It provides a quantitative assessment of the dissimilarity between the generated and real image distributions [23].

4.1. Inpainting Baseline Model

This section compares the performance and time efficiency in achieving the desired results. Specifically, we compare two baseline models: one extracted from the already pre-trained model provided by the GitHub repository [21] and the other that we trained entirely from scratch. By examining Table 1, we observe that both models yielded similar FID scores, indicating comparable image generation quality. However, a noteworthy distinction lies in the significantly reduced training time of our model trained from scratch, as mentioned in previous sections.

Table 1 also reflects the results of another state-of-the-art approach based on Hierarchical VQ-VAE [24]. This method accomplished better results, but it is meant to serve solely as a comparison of some of the best current techniques. Our model does not aim to compete in performance but to contribute to making DMs more efficient.

Table 1. Quantitative evaluation for inpainting on the CelebA-HQ dataset.

Model	FID Score
VQ-VAE [24]	9.78
Pre-trained model	25.55
Ours	24.93

This finding was crucial, as it assured the quality of our model, providing a solid foundation for subsequent experiments involving colorization tasks. By achieving comparable FID scores in a considerably shorter timeframe, we established the viability and efficiency of our training approach, bolstering confidence in the model when we later used it to adapt to colorization functions successfully in subsequent experiments. Moreover, the speed-up in the training process saved valuable time and computational resources and allowed us to iterate more swiftly. This accelerated training pace allowed us to explore a broader range of hyperparameters and even allowed more images and bigger datasets, leading to a deeper understanding of the nuances involved in inpainting tasks. This achievement of comparable FID scores with a significantly expedited training process not only solidifies the foundation of our model but also expands its possibilities, making it a powerful tool for both research and practical applications in the fields of image inpainting and beyond.

On the other hand, Table 2 compares the time required to obtain the mentioned results, highlighting our greatest virtue, the increase in training speed of our model. All training times are associated with DMs. Methods that use other techniques do not apply to this comparison, as we are evaluating our improvement concerning DM performance.

Table 2. Time and efficiency evaluation for inpainting on the CelebA-HQ dataset.

Model	Epochs	Time/Epoch (min)	Total Time (h)
Pre-trained model	200	90	300
Ours	200	8	26.6

Last, a selection of generated images is presented for qualitative evaluation, serving as visual evidence of the performance and quality attained by our model. First, Figure 9 shows the denoising process carried out to perform the inpainting. The top-left image was fed to the model to be denoised. Each of the following figures represents the state at future timesteps until reaching the lower-right image that corresponds to the output and the restored de-noised result. Some more qualitative results are shown in Figures 10 and 11, displayed as pairs of ground truth (left) and generated (right).

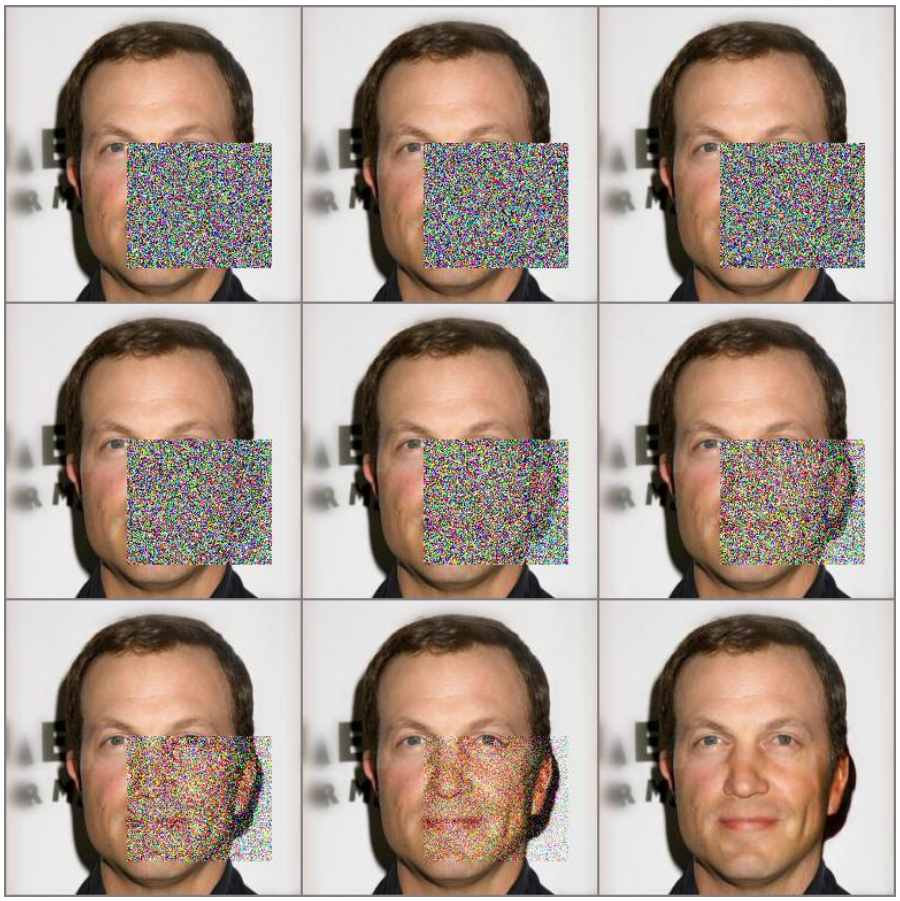

Figure 9. Process of inpainting through denoising.

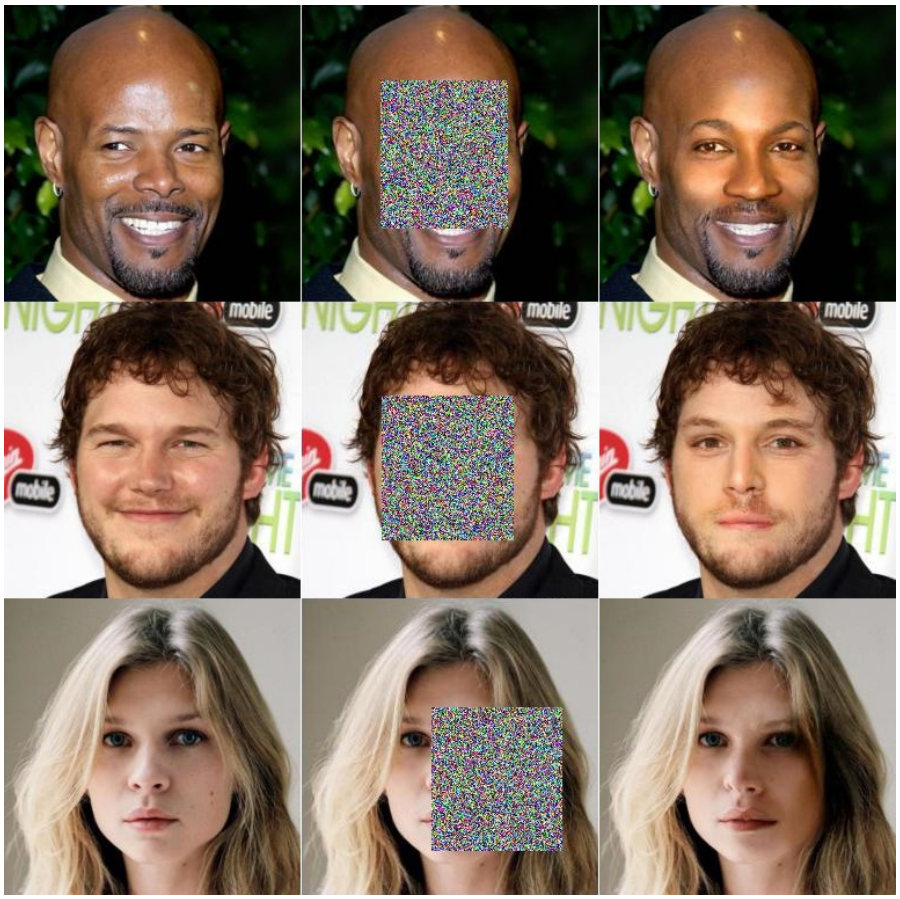

Figure 10. Qualitative comparison for inpainting in the CelebA-HQ dataset. The ground-truth image is on the left, and the generated is on the right. In the center is the masked image.

Figure 11. More qualitative comparison for inpainting in the CelebA-HQ dataset. The ground-truth image is on the left, and the generated is on the right for each pair.

4.2. Face Colorization Model

Now is the opportune moment to assess the performance of our colorization model on the CelebA-HQ faces dataset. Table 3 provides a comprehensive overview of our results compared with a reference approach that utilizes wavelet diffusion models [16] for face colorization purposes.

Table 3. Quantitative evaluation for colorization on the CelebA-HQ dataset.

Model	FID Score
Wavelet DM [16]	6.40
Ours	14.18

One remarkable aspect contributing to our model's efficiency is the small number of training epochs required to achieve these outstanding results. While in the previous section, we observed that the reference approach needed 200 epochs to obtain comparable results and could be further trained, our model achieved this level of performance with just around 20 epochs. This significant reduction in training time speeds up the process immensely and allows us to adapt our model to various tasks within the same domain without requiring extensive training from scratch. Table 4 compares the time required to obtain the mentioned results. For the comparison, we assumed that, if trained from scratch, the model would need another 200 epochs and the time per epoch would be equal to the base model specified in the previous section. On the other hand, our model required much fewer epochs and a much shorter time per epoch. We can see how training a model from scratch that took 26.6 h (as seen in Table 2) allowed us to reduce the training time of obtaining another model with a different scope, only needing 2.6 h to obtain this second one.

Table 4. Time and efficiency evaluation for colorization on the CelebA-HQ dataset.

Model	Epochs	Time/Epoch (min)	Total Time (h)
Standard trained model	200	90	300
Ours	20	8	2.6

This accelerated adaptability of our model opens up numerous beneficial avenues for both research and practical applications in the field. Researchers can now experiment more swiftly with different colorization tasks and fine-tune the model for specific requirements. In practical applications, such as real-time image processing or interactive design tools, this efficiency translates into quicker and more responsive results, greatly enhancing user experiences.

In summary, our colorization model's ability to achieve impressive results with a minimal number of training epochs not only showcases its exceptional performance but also underscores its versatility and efficiency, making it a valuable asset for a wide range of image colorization tasks and applications.

Furthermore, we present visual representations of the generated outputs in Figure 12 for qualitative evaluation. Our colorization model demonstrates impressive performance on the CelebA-HQ faces dataset. The generated results exhibit excellent colorization not only on the facial features but also on the background of the images. It is noteworthy that colorizing the background accurately proved challenging during the initial stages of the training process. However, our model successfully overcame these difficulties and now produces remarkable colorization results across the entire image.

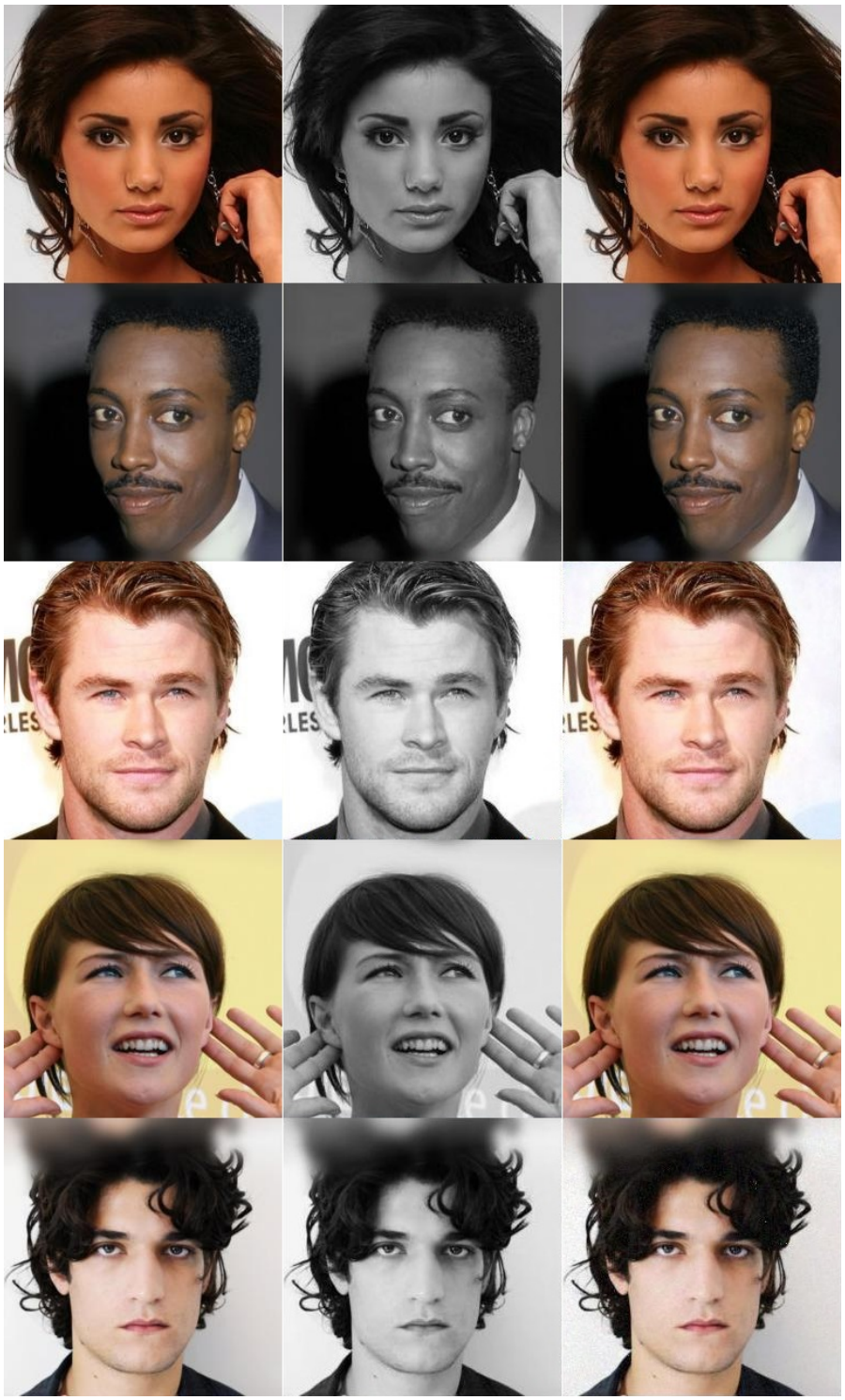

Figure 12. Qualitative comparison for colorization in the CelebA-HQ dataset. The ground-truth image is on the left, and the generated is on the right. In the center is the masked image in a gray-level scale.

4.3. Landscape Colorization Model

Last, we also conducted colorization experiments on the Places-2 dataset, specifically focusing on the forest and mountain subset of images. Our model's performance and a comparison with the Palette approach [1] are summarized in Table 5.

Table 5. Quantitative evaluation for colorization on the Places-2 dataset.

Model	FID Score
Palette [1]	11.70
Ours	17.95

The FID scores indicate the quality of the colorization results, with lower values corresponding to better performance. In this evaluation, the Palette approach achieved a lower FID score of 11.70, while our model obtained a slightly higher score of 17.95. It is important to note that these scores provide a quantitative measure of the performance but may not capture the full qualitative aspects of the colorization results.

Despite the slightly higher FID score, our model still exhibits commendable colorization outcomes for landscape images. The generated results successfully infuse vibrant and accurate colors into the landscapes, enhancing their visual appeal, as seen in Figures 13 and 14. It is worth noting that the two Places-2's dataset selected subsets, although different, hold some overlap, as can be seen in the fourth row of Figure 13, where the lower half can be confused with a forest despite belonging to the class mountain. This selection helps have a wider variety of data and enhances our model, showing its higher capabilities to generalize. Last, there is room for further improvement; our model demonstrates promising capabilities in landscape colorization on the Places-2 dataset. Nevertheless, we present these results for reference purposes.

Additionally, it is worth noting that in this experiment, we also achieved remarkable results with an exceptionally small number of training epochs. In the first section, where we used the reference approach, we trained for 200 epochs to obtain comparable results, and there was still room for further improvement through extended training. However, in this particular experiment, akin to the previous one, we were able to achieve comparable results with just around 20 epochs. This drastic reduction in the required training time signifies a significant leap in efficiency. Table 6 compares the time required to obtain the mentioned results. For the comparison, we assumed once again that, if trained from scratch, the model would need another 200 epochs and the time per epoch would be equal to the base model specified in the previous section. Likewise, our model only needed 20 epochs to adapt from the previous model to this one, maintaining a similar speed-up and increase in efficiency. Being a new dataset, the batches of images fed to the model were slightly bigger, resulting in longer epochs.

Table 6. Time and efficiency evaluation for colorization on the Places-2 dataset.

Model	Epochs	Time/Epoch (min)	Total Time (h)
Standard trained model	200	90	300
Ours	20	10	3.3

What sets this achievement apart is that, unlike the previous scenario, here we focused on adapting the domain of our model while retaining the same task (colorization). In essence, this approach allows us to fine-tune our model for new data distributions. This level of adaptability is particularly valuable in situations where the model needs to adjust to different but related data sources, such as transferring knowledge from one type of image dataset to another while keeping the colorization task consistent.

This demonstration of the model's versatility, which enables it to adapt to new domains and tasks while maintaining its efficiency, underscores the immense potential of diffusion models. While these models are known for their long training times, such as the 200 epochs required for the reference approach, these optimizations and speed-up techniques have now proven effective in drastically reducing this training overhead.

Figure 13. Qualitative comparison for colorization in the Places2 dataset. The ground-truth image is on the left, and the generated is on the right. In the center is the masked image in a gray-level scale.

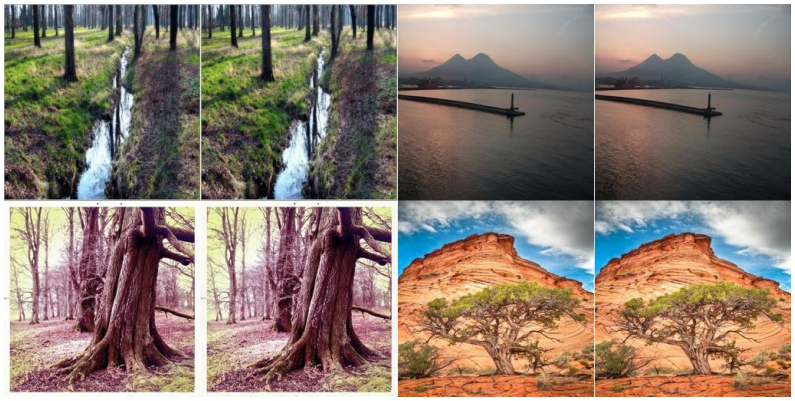

Figure 14. Qualitative comparison for colorization in the Places-2 dataset. The ground-truth image is on the left, and the generated is on the right for each pair.

5. Discussion

The obtained results in our experiments underscore the remarkable performance and efficiency of our model when compared to various baseline approaches. Both the pre-trained model and our model trained from scratch achieved similar FID scores, indicating a comparable level of image generation quality. However, our model truly stands out in

terms of significantly reduced training time. This outcome not only validates the viability but also emphasizes the efficiency of our unique training approach. Solving all three of the models proposed (face inpainting, face colorization, and places colorization) could have taken an unassuming amount of time (900 h), but thanks to our proposal, this could be achieved in 32.6 h (having the greatest overhead obtaining the initial model taking 26.6 h), which led to even further speed-ups the more adaptations we performed. Experiments regarding other domains and/or other tasks adaptation are a matter of future work, but with this experiment as a basis, we can prove the proper functionality of our approach.

It is important to note that while state-of-the-art approaches may achieve slightly better results, they are not direct competitors to our model. Our primary focus revolves around enhancing the efficiency of deep learning models, and the comparable FID scores achieved by our model in a significantly shorter timeframe solidify its quality within this context. Further training and optimizations to achieve state-of-the-art scores are a matter of future work. The qualitative evaluation further reinforces the impressive performance of our model. Visual evidence from our experiments showcases the gradual and effective restoration of images during the inpainting process. Additionally, it highlights the model's exceptional capabilities in colorizing both facial features and complex landscapes. While there may still be room for further refinement and improvement, our model's performance in these tasks is undeniably promising.

Regarding our approach's limitations, there are some aspects to have in mind. The first model, considered the base one, is key for the quality of further adaptations, so beginning with a poorly trained base model can be highly detrimental. Furthermore, another issue our model may suffer from, but one we could not prove, comprises the repercussions and quality loss if too many adaptation steps are applied. In this paper, we reviewed a total of two adaptations, portraying a total of three different models. But, if this number were much higher, some long-training might be needed for some of the steps, especially if the new domains/tasks to adapt to were very different from the original ones.

In summary, the results obtained in our experiments collectively emphasize our model's quality, efficiency, and potential. These findings provide a robust foundation for further exploration and application across a range of tasks and domains. Our model's ability to achieve competitive results with reduced training times makes it a valuable tool for researchers and practitioners seeking efficient and effective deep learning solutions. It also opens up new possibilities for real-world applications where speed and performance are essential, reinforcing its position as a versatile and powerful asset in the field of image processing and beyond.

6. Conclusions

In this study, we address the challenges associated with image-to-image diffusion models in low-resource settings, with a special emphasis on multi-task generalist approaches. While multi-task generalist models offer versatility, they come with computational demands and extended training times that can hinder practicality and accessibility. Building upon the framework established in [1], we propose a novel approach to reduce the computational demands of this kind of multi-task generalist image-to-image diffusion models by applying incremental learning and by leveraging their inherent task/domain transfer learning capabilities.

Our approach starts with a base model trained from scratch for human face inpainting with an incremental learning strategy. The resultant model achieves performance nearly on par with its batch-learning counterpart, with a significant reduction in the training duration. Subsequently, we fine-tune this base model to execute a distinct task, image colorization, demonstrating the significant advantages of adapting this model for other tasks within the same data domain. This adaptation is achieved with remarkable efficiency, reducing training time considerably compared to training from scratch.

Furthermore, we explore the adaptation of a pre-trained model from one domain to another, showcasing the potential to use a single, long-trained model for various tasks

and data domains. This transfer learning approach capitalizes on the knowledge encoded in the pre-trained model, further reducing the computational demands and training time for tackling problems in different domains. We also delve into various configurations, offering insights into the image-to-image diffusion models' potential for cross-task and cross-domain transfer learning.

Our contributions to the field of DMs are two-fold. First, we have developed techniques that enhance the feasibility and accessibility of DMs by reducing their computational requirements and training time. Second, we have demonstrated the adaptability and versatility of DMs, making them valuable tools for image-to-image translation tasks across diverse domains.

In conclusion, this research not only advances the understanding of DMs but also offers practical solutions to overcome their limitations. Our findings provide a foundation for more efficient and accessible image-to-image translation using DMs, paving the way for innovative applications and further exploration in the field. As DMs continue to evolve and adapt, their potential for addressing complex image translation challenges becomes increasingly promising.

Author Contributions: B.O. and R.T.; Methodology, H.A., B.O. and R.T.; Software, H.A.; Validation, B.O. and R.T.; Investigation, H.A.; Writing—original draft, H.A.; Writing—review & editing, B.O. and R.T.; Funding acquisition, B.O. and R.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the Spanish Ministry of Science and Innovation under contracts PID2019-107255GB and PID2021-124463OB-IOO, by the Generalitat de Catalunya under grants 2017-SGR-962, 2021-SGR-00326, 2021-SGR-00478 and DRAC (IU16-011591). The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation program under the HORIZON-AG PHOENI2X (101070586) and HORIZON-EU VITAMIN-V (101093062) projects.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DM	Diffusion Model
GAN	Generative Adversarial Network
FID	Frechet Inception Distance
VLB	Variational Lower Bound

References

1. Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022 Conference, Vancouver, BC, Canada, 7–11 August 2022.
2. Yang, L.; Zhang, Z.; Hong, S. Diffusion models: A comprehensive survey of methods and applications. *arXiv* **2022**, arXiv:2209.00796.
3. Zhang, Q.; Chen, Y. Fast Sampling of Diffusion Models with Exponential Integrator. *arXiv* **2022**, arXiv:2204.13902.
4. Song, Y.; Sohl-Dickstein, J.; Kingma, D.P.; Kumar, A.; Ermon, S.; Poole, B. Score-Based Generative Modeling through Stochastic Differential Equations . In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
5. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
6. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Part III 18; Springer International Publishing: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
7. Kingma, D.; Salimans, T.; Poole, B.; Ho, J. Variational diffusion models. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 21696–21707.

8. Nichol, A.Q.; Dhariwal, P. Improved denoising diffusion probabilistic models. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8162–8171.
9. Gu, S.; Chen, D.; Bao, J.; Wen, F.; Zhang, B.; Chen, D.; Guo, B. Vector quantized diffusion model for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10696–10706.
10. Asperti, A.; Colasuonno, G.; Guerra, A. Portrait Reification with Generative Diffusion Models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; PMLR: Baltimore, MA, USA, 2023.
11. Muhammad, A.; Salman, Z.; Lee, K.; Han, D. Harnessing the power of diffusion models for plant disease image augmentation. *Front. Plant Sci.* **2023**, *14*, 1280496. [CrossRef]
12. Lugmayr, A.; Danelljan, M.; Romero, A.; Yu, F.; Timofte, R.; Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 11461–11471.
13. Zhao, S.; Cui, J.; Sheng, Y.; Dong, Y.; Liang, X.; Chang, E.I.; Xu, Y. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. *arXiv* **2021**, arXiv:2103.10428.
14. Heidari, M.; Morsali, A.; Abedini, T.; Heydarian, S. DiffGANPaint: Fast Inpainting Using Denoising Diffusion GANs. *arXiv* **2023**, arXiv:2311.11469.
15. Zhang, G.; Ji, J.; Zhang, Y.; Yu, M.; Jaakkola, T.; Chang, S. Towards coherent image inpainting using denoising diffusion implicit models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
16. Phung, H.; Dao, Q.; Tran, A. Wavelet diffusion models are fast and scalable image generators. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023.
17. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.
18. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.L.; Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 36479–36494.
19. Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-Or, D.; Bermano, A.H. Human motion diffusion model. *arXiv* **2022**, arXiv:2209.14916.
20. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
21. Palette-Image-to-Image-Diffusion-Models. Available online: https://github.com/Janspiry/Palette-Image-to-Image-Diffusion-Models (accessed on 10 March 2023).
22. Zhou, B.; Khosla, A.; Lapedriza, A.; Torralba, A.; Oliva, A. Places: An image database for deep scene understanding. *arXiv* **2016**, arXiv:1610.02055.
23. Yu, Y.; Zhang, W.; Deng, Y. *Frechet Inception Distance (FID) for Evaluating GANs*; China University of Mining Technology Beijing Graduate School: Beijing, China, 2021.
24. Peng, J.; Liu, D.; Xu, S.; Li, H. Generating diverse structure for image inpainting with hierarchical VQ-VAE. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.