# DECDM: Document Enhancement using Cycle-Consistent Diffusion Models

Jiaxin Zhang Joy Rimchala Lalla Mouatadid Kamalika Das Sricharan Kumar
Intuit AI Research
{jiaxin_zhang, Joy_Rimchala, Lalla_Mouatadid, kamalika_das, sricharan_kumar}@intuit.com

## Abstract

*The performance of optical character recognition (OCR) heavily relies on document image quality, which is crucial for automatic document processing and document intelligence. However, most existing document enhancement methods require supervised data pairs, which raises concerns about data separation and privacy protection, and makes it challenging to adapt these methods to new domain pairs. To address these issues, we propose DECDM, an end-to-end document-level image translation method inspired by recent advances in diffusion models. Our method overcomes the limitations of paired training by independently training the source (noisy input) and target (clean output) models, making it possible to apply domain-specific diffusion models to other pairs. DECDM trains on one dataset at a time, eliminating the need to scan both datasets concurrently, and effectively preserving data privacy from the source or target domain. We also introduce simple data augmentation strategies to improve character-glyph conservation during translation. We compare DECDM with state-of-the-art methods on multiple synthetic data and benchmark datasets, such as document denoising and shadow removal, and demonstrate the superiority of performance quantitatively and qualitatively.*

## 1. Introduction

In our daily lives, we encounter a large number of documents, such as receipts, invoices, and tax forms, that are often degraded in various ways, including noise, blurring, fading, watermarks, shadows, and more, as shown in Figure 1. These degradations can make the documents difficult to read and can significantly impair the performance of OCR systems. Automatic document processing is the first step in document intelligence and aims to enhance document quality using advanced image processing techniques such as denoising, restoration, and deblurring. However, applying these techniques directly to document enhancement may not be effective due to the unique challenges posed by text documents. Unlike typical image restoration tasks, where the degradation function is known and the recovery of the image task can be translated into solving an inverse problem such as inpainting, deblurring/super-resolution, and colorization, real-world document enhancement is a blind denoising process with an unknown degradation function, making it even more challenging. Many state-of-the-art methods have been proposed that rely on assumptions and prior information [15, 34], but there is still a need for more effective techniques that can handle unknown degradation functions.


Figure 1. A performance overview of our DECDM methods on document enhancement tasks, including denoising, shadow removal, binarization, watermark removal, deblur and defade.

Deep learning has led to the development of discriminative models based on convolutional neural networks (CNNs) [45] and auto-encoder (AE) architectures [42], which are important for solving image restorations. However, these methods require noisy/clean paired image data, which is difficult to obtain in real-world applications. Existing benchmark datasets [1] collect clean documents and add synthetic noise, but these do not always accurately represent real-world noise or degradation. To address this, recent works have proposed unpaired ideas based on generative models, such as generative adversarial networks (GANs) [12], which transfer images from one domain to another while preserving content representation [48]. Document denoising can be achieved by transferring from a noisy style to a clean style while preserving the text content. However, these models typically require minimizing an adversarial loss between a specific pair of source and target datasets [29], which has limitations in training instability and potential data privacy leakage [38].

While restoration methods have shown their capability of producing high-quality restorations, they have a severe limitation in their adaptability to different domains [11]. These

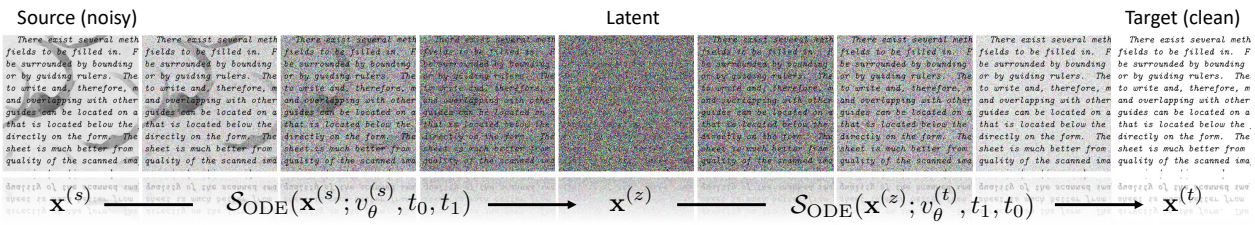


Figure 2. Cycle-Consistent Diffusion Models leverages two deterministic diffusions through ODEs for unpaired document-level image-to-image translation. Given source data $\mathbf{x}^{(s)}$, the source diffusion model $v_\theta^{(s)}$ runs in the forward direction to convert it to the latent space $\mathbf{x}^{(z)}$, while the target diffusion model $v_\theta^{(t)}$ reverse ODE to construct the target document-level images $\mathbf{x}^{(t)}$. $t_0$ and $t_1$ are the starting point and ending point, typically setting to $t_0 = 0$ and $t_1 = 1$.

models are trained on specific noise-clean pairs, making it difficult to use them for restoration tasks in different domains. For instance, a model designed for watermark removal may not perform well for denoising tasks. This domain-specific training leads to a significant increase in the number of models required for different domain pairs, making it computationally prohibitive. Additionally, these models are trained using joint data from both source and target domains, thereby raising concerns over *data privacy protection*. This issue may be critical in certain privacy-sensitive applications, such as financial documents and medical imaging, where data providers may be reluctant to share their data.

Beyond both disadvantages of existing methods, the task of document enhancement presents several unique challenges compared to typical image translation problems. These include (1) *High-resolution*, which poses scalability challenges, leading to performance degradation and significant increases in training costs. (2) *Lack of large benchmark datasets*, which makes it infeasible to use large pre-trained models. While the success of large generative models such as Stable diffusion [24], Dall·E [23], and Imagen [26] is largely attributed to large datasets, such as LAION-5B [28], there is currently no large pre-trained model available for document-level tasks. (3) *Character feature damage*. Unlike image translation at the pixel level, document-level image translation requires preserving original content such as characters and words while accounting for style differences in the background, i.e., noise to clean. Current methods only focus on pixel-level information and do not consider critical character features such as glyphs, resulting in character-glyph damage during the translation process [30].

In this work, we present DECDM, an unsupervised end-to-end document-level image translation method that addresses the challenges faced by existing document enhancement methods. Inspired by recent advances in diffusion models [32, 34, 38, 41], our approach independently trains the source (noisy) and target (clean) models, decoupling paired training and enabling the domain-specific diffusion models to remain applicable to other pairs. Specifically, we build DECDM based on denoising diffusion implicit models (DDIMs) [32], which create a deterministic and reversible mapping between images and their latent representations, solved using ordinary differential equation (ODE) that forms the cornerstone. Translation with DECDM on a source-target pair requires two different ODEs: the source ODE encodes input images to the latent space, while the target ODE decodes images in the target domain, as shown in Figure 2.

Since training diffusion models are specific to individual domains and rely on no domain pair information, DECDM makes it possible to save a trained model of a certain domain for future use, when it arises as the source or target in a new pair. Pairwise translation with DECDM requires only a linear number of diffusion models, which can be further reduced with conditional models [9]. Additionally, the training process focuses on one dataset at a time and does not require scanning both datasets concurrently, preserving the data privacy of the source or target domain.

To overcome the challenges in document-level translation, we propose a simple data augmentation scheme to downscale the resolution of training data, while significantly increasing the dataset size. This approach reduces the diffusion training cost and improves the performance in learning character distribution benefiting from large datasets. Experimentally, we demonstrate the effectiveness of DECDM on a variety of document enhancement tasks, such as document denoising and document shadow removal, with qualitative and quantitative results that establish DECDM as a scalable, efficient, and reliable solution to the family of document enhancement approaches. DECEM is also well-suited for few-shot scenarios by leveraging unpaired training and sample efficiency in cycle-consistent diffusion models and data augmentation strategies. Beyond the denoising and removal tasks shown here, our proposed DECDM method can apply to broader few-shot document enhancement tasks in Figure 1.

## 2. DECDM Method

Our goal is to develop a cycle-consistent diffusion model for document enhancement by solving the following three core problems: (1) unpaired supervision, (2) enforcing cycle consistency, and (3) data privacy protection. Then we intro-

| Methods | Unpaired or paired | Backbone Models | | | Document Enhancement Tasks | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | GANs | CNNs | Transformers | Denoise | Shadow Removal | Binarization | Watermark Removal | Deblur | Defade |
| SCGAN [43] (ICCV 17') | Paired | ✓ | - | - | - | - | - | - | ✓ | - |
| SCDCA [46] (ICPR 18') | Paired | - | ✓ | - | ✓ | - | - | - | ✓ | - |
| BEDSR-Net [20] (CVPR 20') | Paired | ✓ | - | - | - | ✓ | - | - | - | - |
| DE-GAN [37] (TPAMI 20') | Paired | ✓ | - | - | - | - | - | ✓ | ✓ | - |
| RED-Net [4] (PR 19') | Paired | - | ✓ | - | - | - | ✓ | - | - | - |
| SauvolaNet [18] (ICDAR 21') | Paired | - | ✓ | - | - | - | ✓ | - | - | - |
| CharFormer [30] (ACM MM 22') | Paired | - | - | ✓ | ✓ | - | - | - | - | - |
| DocEnTr [36] (ICPR '22) | Paired | - | - | ✓ | - | - | ✓ | - | ✓ | ✓ |
| CycleGAN [29] (ACCV 18') | Unpaired | ✓ | - | - | - | - | - | ✓ | ✓ | ✓ |
| CycleGAN-MOE [11] (ICCV 21') | Unpaired | ✓ | - | - | ✓ | - | - | ✓ | ✓ | ✓ |

Table 1. A summary of document enhancement methods, including unpaired/paired supervision, backbone models (CNNs, GANs, Transformers), and enhancement tasks (denoise, shadow removal, binarization, watermark removal, deblur, defade).

duce the data augmentation strategies for dealing with the challenges of document datasets while improving character and word feature preservation.

## 2.1. Problem Formulation

We first define the unpaired document enhancement task from a mathematical perspective as follows:

**Problem 1** (Unpaired Document Enhancement). *Given two unpaired sets of documents, one set consisting of degraded documents $\mathcal{X}$ (source domain), and the other a collection of clean documents $\mathcal{Y}$ (target domain), our goal is to learn a mapping $\mathcal{F}: \mathcal{X} \rightarrow \mathcal{Y}$ such that the output $\hat{y} = \mathcal{F}(\mathbf{x}), \mathbf{x} \in \mathcal{X}$, is indistinguishable from documents $\mathbf{y} \in \mathcal{Y}$ to classify $\hat{\mathbf{y}}$ apart from $\mathbf{y}$.*

The degraded documents include multiple types, e.g., noise, blurring, watermark, etc, as shown in Fig. 1. The mapping $\mathcal{F}$ should satisfy two conditions: content preservation and style transfer. The content refers to the character, text, numbers, tables, and figures in documents and the style transfer means the translation from degraded documents (source domain $\mathcal{X}$) to clean documents (target domain $\mathcal{Y}$). Our objective is therefore to convert the degraded documents in $\mathcal{X}$ while preserving their core contents in $\mathcal{Y}$. From the computer vision perspective, enhancement tasks can be essentially interpreted as document-level image-to-image translation.

**Problem 2** (Cycle Consistency). *Assuming we have a mapping $\mathcal{F}: \mathcal{X} \rightarrow Y$ and another mapping $\mathcal{H}: \mathcal{Y} \rightarrow \mathcal{X}$, then $\mathcal{F}$ and $\mathcal{H}$ should be inverse of each other, and both mappings should be bijective, i.e., satisfying*

$$\mathcal{F}(\mathcal{H}(\mathbf{x})) \approx \mathbf{x}, \quad \mathcal{H}(\mathcal{F}(\mathbf{y})) \approx \mathbf{y} \tag{1}$$

A desirable feature of image translation algorithms is the *cycle consistency* property [48], which transforms a sample in the source domain to the target domain, and then back to the source, will recover the original sample in the source domain. This property is critical to the adaptability guarantee, which empowers the domain-specific diffusion models to stay applicable in other pairs. A rigorous formulation is defined in Eq. (1).

**Problem 3** (Data Privacy). *In the training and translation process, source model $v_\theta^{(s)}$ and target model $v_\theta^{(t)}$ are decoupled and trained independently, while both source datasets $\mathbf{x} \in \mathcal{X}$ and target datasets $\mathbf{y} \in \mathcal{Y}$ are private to each other.*

Most image-to-image translation approaches strongly rely on joint training over data from both source domains and target domains. This leads to a significant challenge in preserving the privacy of domain data in a federated setting. An ideal method is to train the models independently on separate domain datasets such that data privacy is protected.

## 2.2. Cycle-Consistent Diffusion Models

Diffusion Models [13, 31, 33] aim at modeling a distribution $p_\theta(\mathbf{x}_0)$ to approximate the data distribution $q(\mathbf{x}_0)$ through diffusion and reversed generative processes. Song et al. [35] proposed a unified framework by leveraging Stochastic Differential Equations (SDEs) representation, which uses a forward and backward SDE to mathematically describe general diffusion processes:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w} \tag{2}$$

and reversed generative processes:

$$\mathrm{d}\mathbf{x} = [\mathbf{f} - g^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})]\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w} \tag{3}$$

where $\mathbf{f}(\mathbf{x}, t)$ is the vector-valued coefficient, $\mathbf{w}$ is the standard Wiener process, $g(t)$ is the diffusion coefficient, and $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the score function of the noise perturbed data distribution. Any diffusion process can be represented by a deterministic ODE [35], named the probability flow (PF) ODE [35], which enables uniquely identifiable encodings of data, and has the following form:

$$\mathrm{d}\mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - \frac{1}{2} g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right]\,\mathrm{d}t \tag{4}$$

which is equivalent to the forward SDE in Eq. (2). For conciseness, we use $\theta$-parameterized score networks $\mathbf{s}_{t,\theta} \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ to approximate the score function and use $v_\theta = \mathrm{d}\mathbf{x}/\mathrm{d}t$ to denote the $\theta$-parameterized model and use the symbol $\mathcal{S}_{\text{ODE}}$ to denote the mapping from $\mathbf{x}^{(t_0)}$ to $\mathbf{x}^{(t_1)}$ and implement ODE solver in DDIMs [32].

$$\begin{aligned}\mathbf{x}(t_1) &= \mathcal{S}_{\text{ODE}}(\mathbf{x}(t_0); v_\theta, t_0, t_1) \\ &= \mathbf{x}(t_0) + \int_{t_0}^{t_1} v_\theta(t, \mathbf{x}(t))\, \mathrm{d}t\end{aligned} \tag{5}$$

In this work, we implement an ODE solver in DDIMs [32] where the generative sampling process is defined in a deterministic non-Markovian manner, which can be used for the reverse direction, deterministically noising an image to obtain the initial noise vector. This property is central to DECDM as we solve these ODEs for forward and reverse conversion between data and their latents.

**Cycle-Consistent Diffusion Models.** DECDM leverages the cycle-consistent diffusion models to perform unpaired document-level image translation, with two diffusion models trained independently on two separate domains. DECDM consists of two core steps, training and translation, described in Algorithms 1 and 2. For training, DECDM first collects noisy data from the source domain $\mathbf{x}^{(s)} \sim p_s(\mathbf{x})$, and clean data from the target domain $\mathbf{x}^{(t)} \sim p_t(\mathbf{x})$, then train two diffusion models separately on the two domains and save them as $v_\theta^{(s)}$ and $v_\theta^{(t)}$. For translation, DECDM first runs $\mathcal{S}_{\text{ODE}}$ in the source domain to obtain the latent encoding $\mathbf{x}^{(z)}$ of the image $\mathbf{x}^{(s)}$ at the end time $t_1$ via $\mathcal{S}_{\text{ODE}}(\mathbf{x}^{(s)}; v_\theta^{(s)}, t_0, t_1)$. Then DECDM feds the source latent encoding $\mathbf{x}^{(z)}$ to $\mathcal{S}_{\text{ODE}}$ with the target model $v_\theta^{(t)}$ to reconstruct the target image $\mathbf{x}^{(t)}$ via $\mathcal{S}_{\text{ODE}}(\mathbf{x}^{(z)}; v_\theta^{(t)}, t_1, t_0)$, as illustrated in Figure 2.

One of the important advantages of DECDM is the exact cycle consistency: transforms a sample in the domain $\mathcal{S}$ to the domain $\mathcal{T}$, and then back to $\mathcal{S}$, will recover the original sample in $\mathcal{S}$. As probability flow ODEs are used, the cycle consistency property is guaranteed [35]. The following proposition validates the cycle consistency of DECDM.

**Proposition 4** (Exact Cycle Consistency). *Given a specific sample $\mathbf{x}^{(s)}$ from source domain $\mathcal{X}$, with a trained source model $v_\theta^{(s)}$ and a target model $v_\theta^{(s)}$, we define the forward cycle consistency*

$$\begin{aligned}\mathbf{x}^{(z)} &= \mathcal{S}_{\text{ODE}}(\mathbf{x}^{(s)}; v_\theta^{(s)}, t_0, t_1); \\ \mathbf{x}^{(t)} &= \mathcal{S}_{\text{ODE}}(\mathbf{x}^{(z)}; v_\theta^{(t)}, t_1, t_0);\end{aligned} \tag{6}$$

*and backward cycle consistency*

$$\begin{aligned}\tilde{\mathbf{x}}^{(z)} &= \mathcal{S}_{\text{ODE}}(\mathbf{x}^{(t)}; v_\theta^{(t)}, t_0, t_1); \\ \tilde{\mathbf{x}}^{(s)} &= \mathcal{S}_{\text{ODE}}(\tilde{\mathbf{x}}^{(z)}; v_\theta^{(s)}, t_1, t_0);\end{aligned} \tag{7}$$

*Assume zero discretization error, then we have $\mathbf{x}^{(s)} = \tilde{\mathbf{x}}^{(s)}$.*

In practice, we implement the ODE solver $\mathcal{S}_{\text{ODE}}$ with DDIMs [32] which has reasonably small discretization errors. Thus DECDM incurs almost negligible cycle inconsistency.

## 2.3. Data Privacy Protection

The DECDM training process does not depend on knowledge of the domain pair a priori, while only source and target data are required. Both source and target diffusion models are trained independently. The DECDM translation process can be performed in a privacy-sensitive manner. For example, user A is the owner of the source domain and user B is the owner of the target domain. User A intends to translate the source images to the target domain in a private manner without releasing the source dataset. User B also wishes to make the target dataset private. In such a case, user Acan can simply train a diffusion model with the source data, encode the data to the latent space, and only transmit the latent codes to user B. Then user B can use the pretrained diffusion models (using the target data) to convert the received latent code to a target image and send back to user A. The process only requires shared latent code from user A and a pretrained model from user B, which can be finished in a private platform, and both source and target datasets are private to the two parties. This is a significant advantage of DECDM over alternate methods, as we enable strong privacy protection of the datasets. More discussions can be found in the supplementary material.

## 2.4. Data Augmentation

Many document benchmark datasets are not large enough for diffusion model training such that data augmentation is often necessary. However, typical image data augmentation techniques, e.g., crop, rotate, flip, etc, may negatively affect the recognition (difficult to read) of character and word contents. In this work, we implement two simple strategies for document-level data augmentation, while mitigating the high-resolution challenges such as computational scalability issues in training diffusion models, as shown in Fig.4.

The sub-window strategy divides the high-resolution images into several smaller domains, e.g., 1024x1024 images will be divided into 16 sub-images (256×256) or 64 sub-images (128×128). Using this way, we reduce the image resolution but upscale the dataset size fed to the diffusion models for better performance at a lower training cost. If the data is very sparse, we can consider the slide-window strategy, which is inspired by convolution operation in CNN, moving the sub-window with a specific stride. This strategy will significantly increase the amount of data which allows diffusion models to accurately capture the distribution of characters and words. For translation, we perform the same strategy for the source (noisy) data and obtain the corre-

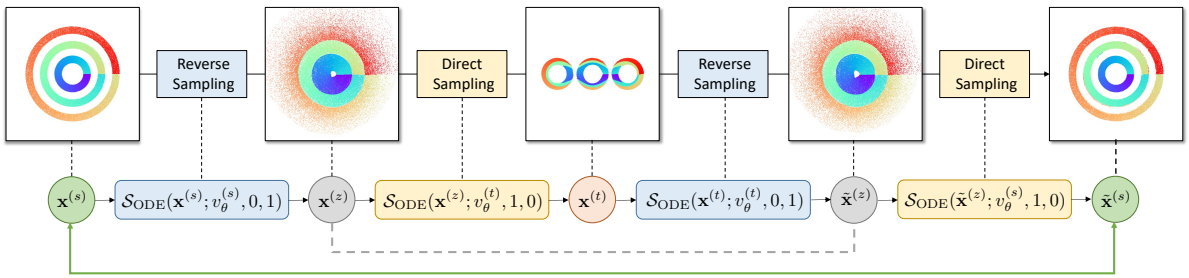


Figure 3. Cycle consistency illustration. Translation from the source domain (CR) to the target domain (PR) and then back to the source domain (CR) via the cycle-consistent diffusion models with reverse and direct sampling.

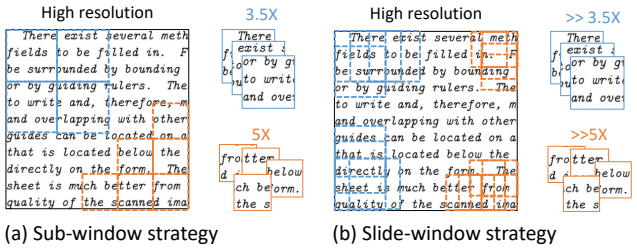


Figure 4. Data argumentation for document-level high-resolution images: (a) sub-window strategy and (b) slide-window strategy.

sponding target sub-images, and finally we ensemble all of them to obtain the whole cleaned images.

**Algorithm 1** Diffusion model training in DECDM

1: **Requirement**: noise data from source domain, $\mathbf{x}^{(s)} \sim p_s(\mathbf{x})$, clean data from target domain, $\mathbf{x}^{(t)} \sim p_t(\mathbf{x})$.
2: Perform data augmentation for $\mathbf{x}^{(s)}$ and $\mathbf{x}^{(t)}$
3: Train source diffusion model $v_\theta^{(s)}(\mathbf{x}^{(s)}) \approx p_s(\mathbf{x})$ and target diffusion model $v_\theta^{(t)}(\mathbf{x}^{(t)}) \approx p_t(\mathbf{x})$ separately
4: **Return** trained source model $v_\theta^{(s)}$ and target model $v_\theta^{(t)}$

**Algorithm 2** Unpaired image translation in DECDM

1: **Requirement**: data sample from source domain $\mathbf{x}^{(s)} \sim p_s(\mathbf{x})$, source model $v_\theta^{(s)}$, target model, $v_\theta^{(t)}$, $t_0, t_1$
2: **Encoding**: obtain latent embedding from source domain data via $\mathbf{x}^{(z)} = \mathcal{S}_{\text{ODE}}(\mathbf{x}^{(s)}; v_\theta^{(s)}, t_0, t_1)$;
3: **Decoding**: obtain target domain data reconstructed from latent code via $\mathbf{x}^{(t)} = \mathcal{S}_{\text{ODE}}(\mathbf{x}^{(z)}; v_\theta^{(t)}, t_1, t_0)$
4: **Return**: $\mathbf{x}^{(t)}$

## 3. Experiments

A set of experiments are provided to demonstrate the effectiveness of our DECDM. We first use a 2D synthetic example to show the cycle-consistent property and then demonstrate DECDM on various document enhancement tasks, including dirty document denoising and shadow removal.

### 3.1. 2D Synthesis Examples

We perform domain distribution translation on two-dimensional synthetic datasets with complex shapes and configurations, as shown in Fig. 5. In this example, we use six 2D datasets (normalized to zero mean and identify covariance): Two Moons (TM); Checkerboards (CB); Concentric Rings (CR); Concentric Squares (CS); Parallel Rings (PR); and Parallel Squares (PS). The colors in Fig. 5 are signed based on the point identities that can help check if a point in the source domain is blue, then its corresponding point in the target domain is also colored blue. To this end, we observed a smooth translation between the source and target domain with point identity preservation. For instance, on the second row in Fig. 5, the red points in the CR dataset are mapped to similar coordinates (relative location) in the target domain of the CS dataset. The latent space provides a disentangled representation of this domain translation.

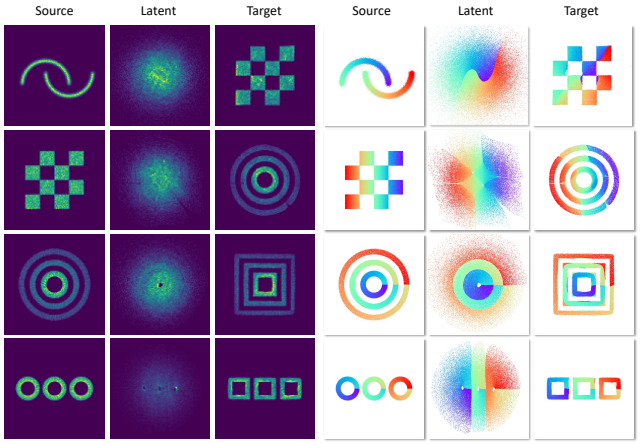


Figure 5. Distribution translation of synthetic datasets: from source datasets to latent representation via encoding, then from latent representation to target datasets via decoding. (Left three) heatmap results and (Right three) scatter results with color configurations.

**Cycle Consistency Validation**. We demonstrate the cycle consistency using an example of domain translation from CR to PR, as shown in Fig. 3. We first train the cycle-consistent diffusion models for each domain (CR and PR) indepen-

dently. Then starting from the CR dataset $\mathbf{x}^{(s)}$, we obtain the latent points $\mathbf{x}^{(z)}$ using reverse sampling and construct the target PR points $\mathbf{x}^{(t)}$ via direct sampling. The next step is the reverse direction, i.e., transforming the target PR points back to the latent and the source CR domain. Similarly, we transfer $\mathbf{x}^{(t)}$ to the latent points $\tilde{\mathbf{x}}^{(z)}$ using reverse sampling and then reconstruct the source CR domain $\tilde{\mathbf{x}}^{(s)}$ via direct sampling. After this multi-step trip, the source points are approximately mapped back to their original positions. From Fig. 3, we observed a similar color topology both in the latent and source domain. The reconstructed source points $\tilde{\mathbf{x}}^{(s)}$ are highly consistent with the original source points $\mathbf{x}^{(s)}$. To further compare the difference, Table 2 shows quantitative evaluation results on cycle consistency among various cases. We use averaged L2 distance to measure the difference between the original points and the reconstructed points after cycle translation, e.g., "TM-CB" means TM $\rightarrow$ CB $\rightarrow$ TM. The results in Table 2 are negligibly small in terms of both the latent and source domains such that the cycle consistency is valid even without adding cycle-consistent loss [48].

| Distance | TM-CB | CR-TB | CR-CS | CR-PR | PR-PS | PS-CS |
|---|---|---|---|---|---|---|
| Latent | 0.0128 | 0.0087 | 0.0101 | 0.0120 | 0.0092 | 0.0100 |
| Source | 0.0122 | 0.0106 | 0.0082 | 0.0108 | 0.0143 | 0.0065 |

Table 2. Cycle consistency validation. Averaged L2 distance is used to measure the difference between original points and after cycle translation on both latent and source domains.

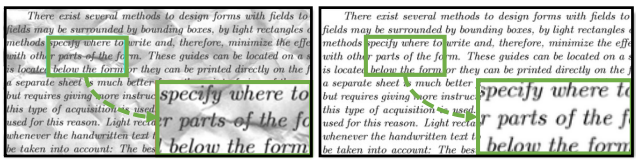


Figure 6. Visualization of DatasetA: (Left) raw document-level image and (Right) ground truth, which is the clean image.

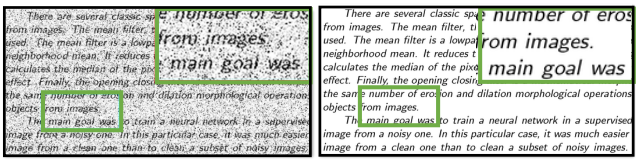


Figure 7. Visualization of DatasetB: (Left) noisy document-level image and (Right) ground truth.

### 3.2. Dirty Document Denoising

**Datasets**. In this case, we apply our DECDM for denoising dirty documents by leveraging the benchmark datasets *denoising-dirty-document*[1], which consists of printed English words in 18 different fonts. The original datasets include noisy raw document-level images with uneven backgrounds, e.g., watermarks, messy artifacts, etc. We name the original datasets as *DatasetA: Dirty Document*. There are 144 data for training and 72 data for testing in the original setting. We use this setting for evaluating all the methods. To increase the complexity, we also create *DatasetB: Noisy Document* by adding speckle noise and Gaussian noise on the ground truth. The noise means $\mu$ is 0 and variance $\sigma$ is 5, which follows the setting in [30]. Fig. 6 shows one of the raw document-level images and the corresponding clean image in DatasetA. Fig. 7 shows the noisy document-level image in DatasetB.

**Baselines**. We compare our DECDM with multiple competitive baseline methods, including GAN/CNN-based methods, CIDG [44], InvDN [21], CycleGAN [29], and some Transformer-based methods, i.e., UFormer [40], IPT [5], TransUNet [6] and CharFormer [30]. Note that most of these state-of-the-art methods are proposed for general image denoising or restoration, not specifically designed for document denoising. Thus, we use the same training environment and datasets for all the methods and report the results if they have already been provided in their work [30]. We perform a slide-window strategy for data augmentation in this case and all the experiments and comparisons are done on one NIVIDA Tesla V100 GPU.

| Method | DatasetA | | | DatasetB | | |
|---|---|---|---|---|---|---|
| | PSNR↑ | SSIM↑ | AC↑ | PSNR↑ | SSIM↑ | AC↑ |
| Raw Data | 16.33 | 0.7978 | 0.6931 | 13.03 | 0.2852 | - |
| CIDG [44] | 21.88 | 0.8871 | 0.7559 | 20.65 | 0.8623 | 0.2471 |
| InvDN [21] | 22.40 | 0.8807 | 0.8374 | 20.49 | 0.8077 | 0.5917 |
| CycleGAN [29] | 23.66 | 0.8857 | 0.8319 | 20.97 | 0.8470 | 0.6409 |
| UFormer [40] | 23.86 | 0.8970 | 0.8326 | 21.01 | 0.8221 | 0.6693 |
| IPT [5] | 23.72 | **0.9027** | 0.856 | 21.94 | 0.8293 | 0.6854 |
| TransUNet [6] | 23.92 | 0.8998 | **0.8621** | 20.83 | 0.8592 | 0.5579 |
| CharFormer [30] | **24.08** | 0.8985 | 0.8553 | **21.07** | **0.8637** | **0.7259** |
| DECDM | **24.30** | **0.9058** | **0.8714** | **21.12** | **0.8631** | **0.7438** |

Table 3. Quantitative evaluation results on average PSNR, SSIM and OCR accuracy (AC). The best two results are highlighted in bold black.

**Metrics**. We introduce two commonly used metrics to evaluate the document-level denoising performance, i.e., peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM). Note that "↑" represents the higher the metric the higher image quality. Additionally, we introduce a metric for evaluating the character-level quality, i.e., optical character recognition (OCR) accuracy (AC). This metric allows us to validate if the denoising algorithms improve the OCR[2] performance compared to dirty documents.

**Qualitative Evaluation**. We first visualize the denoising results by using DECDM and compare it with other baseline methods. Fig. 8 and Fig. 9 show the qualitative performance on *DatasetA* and *DatasetB* respectively. DECDM can effectively remove messy dirties and even backgrounds and

[1]https://www.kaggle.com/competitions/denoising-dirty-documents

[2]The public OCR tools can be accessed via https://www.ocr2edit.com

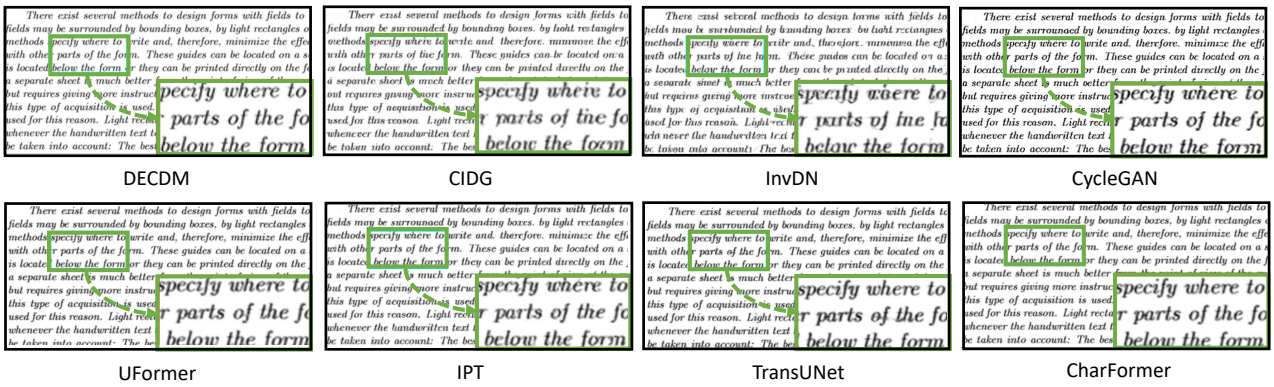


Figure 8. Qualitative evaluations and comparisons on *DatasetA* which is dirty document denoising.

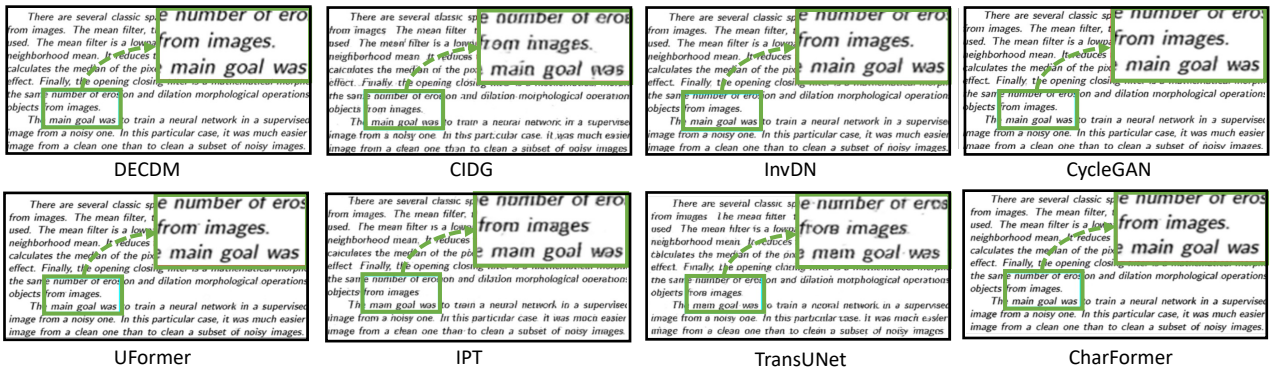


Figure 9. Qualitative evaluations and comparisons on *DatasetB* which is dirty document denoising.

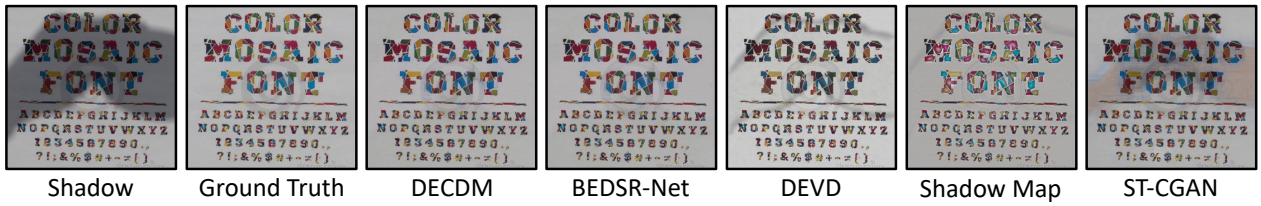


Figure 10. Qualitative evaluation and visual comparison of competing baseline methods on document shadow removal task.

| Method | SDSRD [20] | | RDSRD [20] | | SM Datasets [2] | | DVED Datasets [16] | | WF Datasets [14] | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ | PSNR ↑ | SSIM ↑ |
| Raw Shadow Images | 22.80 | 0.8992 | 21.73 | 0.8093 | 28.45 | 0.9742 | 19.31 | 0.8429 | 20.35 | 0.8850 |
| Shadow Map [2] | 31.55 | 0.9658 | 28.24 | 0.8664 | **35.22** | **0.9823** | 29.66 | 0.9051 | 23.70 | 0.9015 |
| DVED [16] | 22.03 | 0.8435 | 22.53 | 0.7056 | 26.50 | 0.8381 | 26.45 | 0.8481 | 24.45 | 0.8332 |
| Water Filling [14] | 17.06 | 0.8226 | 14.45 | 0.7054 | 13.88 | 0.8059 | 19.21 | 0.8724 | **28.49** | 0.9108 |
| ST-CGAN [39] | 39.38 | 0.9834 | 30.31 | 0.9016 | 29.12 | 0.9600 | 25.92 | 0.9062 | 23.71 | 0.9046 |
| BEDSR-Net [20] | **43.59** | **0.9935** | **33.48** | **0.9084** | **35.07** | **0.9809** | **32.90** | **0.9354** | 27.23 | **0.9115** |
| DECDM | **45.73** | **0.9932** | **37.21** | **0.9143** | 34.95 | 0.9642 | **35.01** | **0.9521** | **29.87** | **0.9112** |

Table 4. Quantitative evaluation results on PSNR and SSIM. We compare our DECDM with BEDSR-Net [20], ST-CGAN [39], Water Filling [14], DVED [16], and Shadow Map [2] methods. The best two results are highlighted in black bold.

perform high-quality document-level image denoising. Unlike some methods, e.g., CycleGAN, InvDN, and TransUNet with character-level damages, DECDM well recognizes the character style and topology, which can be clearly seen in the zoom-in sub-figures in Fig. 8. As an unpaired method, DECDM shows competitive performance compared to the transformer-based methods, e.g., CharFormer and UFormer, which strongly rely on paired supervision. More ablation studies are provided in the supplementary material.

**Quantitative Evaluation**. Table 3 shows the quantitative comparisons between DECDM and state-of-the-art baseline methods on both datasets. Clearly, DECDM shows outperformed results, specifically the AC metric, in both datasets. Compared with GAN/INN models, transformer-based models perform competitively, e.g., CharFormer in DatasetB but it will fail in the unsupervised setting.

### 3.3. Document Shadow Removal

**Datasets**. Although there exist a few datasets for document image shadow removal, they are only used for evaluation on a small scale. In this example, we consider the following five datasets ranging from small-scale to large-scale such that we can provide a comprehensive validation.

- SDSRD datasets [8,20]: 8309 paired images from 970 documents, including synthetic, diverse contexts and lighting. 7533 for training and 776 for testing.
- RDSRD datasets [20]: 540 paired images of 25 documents, including newspaper, slides, and paper, under different lighting conditions.
- Shadow Map (SM) datasets [2]: 81 paired images with light shadows/text only.
- DEVD datasets [16]: 300 paired document-level images, including dark shadows and colorful symbols.
- Water-Filling datasets [14]: 87 high-quality paired images including multi-cast shadows.

**Baselines**. We compared our DECDM with five state-of-the-art methods, including BEDSR-Net [20], ST-CGAN [39], Water Filling [14], DVED [16], and Shadow Map [2] methods. For a fair comparison, we used the publicly available source codes or reported results provided by the authors. We evaluate the compared methods from visual quality using the PSNR and SSIM metrics, as suggested by [20].

**Qualitative and Quantitative Evaluation**. For visual comparison, Fig. 10 shows several shadow removal results of the compared methods. DEVD [16] and ST-CGAN [39] exhibit remaining shadow edges and Shadow Map [2] performs better than those two but still shows the shadow. DECDM close to BEDSR-Net [20] shows ideal performance without seeing shadow edges. Quantitatively, DECDM outperforms other baselines on most datasets as shown in Table 4. For SM datasets, Shadow Map performs best but its result is worse than the other baselines in the other four datasets. BEDSR-Net is a competitive method that achieves promising results but it strongly relies on the pair datasets. On the contrary, DECDM is more flexible and robust without the assumption of pair knowledge such that we can easily deploy it in more real-world scenarios. We also provide a detailed analysis of the effect of data augmentation strategies in the supplementary material.

## 4. Related Work

**Document Enhancement.** Deep learning has enabled many approaches for enhancing the quality of document-level images [1]. Recent state-of-the-art methods in document enhancement are summarized in Table 1, categorized by their supervision mechanism (paired or unpaired), backbone models (CNNs [4,19,46], GANs [11,20,29,37,43], and Transformers [30,36]), and enhancement tasks (denoising, shadow removal, binarization, watermark removal, deblur, and defade). Although most methods perform well in one or multiple tasks, no single model can handle all types. Additionally, paired supervision is required, which is rarely met in real settings. While Cycle-GAN [11, 29] methods can mitigate this limitation, they still need to optimize for cycle consistency over two domains, leading to instability issues and potential data privacy leakage. Our proposed DECDM addresses these challenges by enabling unpaired translation, cycle consistency, and data privacy protection.

**Diffusion Models.** Diffusion models are a family of generative models that have gained much attention recently due to their superior performance in text-guided image synthesis [3, 10, 25], e.g., Stable Diffusion [24], DALL·E 2 [23], and Imagen [26]. These works are built upon the foundation of diffusion models, including score-based methods [33,35] that match with Langevin dynamics, denoising diffusion probabilistic models (DDPMs) [13,31] that parameterize the ELBO objective with Gaussian, and denoising diffusion implicit models (DDIMs) [32] that accelerate DDPM inference via non-Markovian processes. Recent works have leveraged diffusion models for image editing [7,17,27,41], composition [22,47], and restoration tasks [15,26] with promising performance. However, these methods mostly relied on joint training by leveraging both datasets directly. Our DECDM performs a decoupled mechanism by applying separate, pre-trained diffusion models and leveraging the geometry of the shared space for document image translation. To the best of our knowledge, DECDM is the first work to apply diffusion models for document enhancement via unpaired image translation, inspired by these studies.

## 5. Conclusions

DECDM provides an unsupervised end-to-end solution for document image enhancement that offers several advantages over existing state-of-the-art methods, including adaptability to new domain pairs and data privacy protection. These unique capabilities make DECDM a more robust, safe, and scalable solution for improving OCR performance in a wide range of document enhancement tasks. Future works aim to address the current limitations caused by data sparsity, augmentation, and character/word context recognition. We will also integrate OCR into the training pipeline to pursue better character and word recognition.

## References


[1] Zahra Anvari and Vassilis Athitsos. A survey on deep learning based document image enhancement. *arXiv preprint arXiv:2112.02719*, 2021. 1, 8

[2] Steve Bako, Soheil Darabi, Eli Shechtman, Jue Wang, Kalyan Sunkavalli, and Pradeep Sen. Removing shadows from images of documents. In *Asian Conference on Computer Vision*, pages 173–183. Springer, 2016. 7, 8

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 8

[4] Jorge Calvo-Zaragoza and Antonio-Javier Gallego. A selectional auto-encoder approach for document image binarization. *Pattern Recognition*, 86:37–47, 2019. 3, 8

[5] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 6

[6] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6

[7] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. Ilvr: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356. IEEE Computer Society, 2021. 8

[8] Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. Icdar2017 competition on recognition of documents with complex layouts-rdcl2017. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1404–1410. IEEE, 2017. 8

[9] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[10] 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. *arXiv preprint arXiv:2212.05032*, 2022. 8

[11] Mehrdad J Gangeh, Marcin Plata, Hamid R Motahari Nezhad, and Nigel P Duffy. End-to-end unsupervised document image blind denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7888–7897, 2021. 1, 3, 8

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3, 8

[14] Seungjun Jung, Muhammad Abul Hasan, and Changick Kim. Water-filling: An efficient algorithm for digitized document shadow removal. In *Asian Conference on Computer Vision*, pages 398–414. Springer, 2018. 7, 8

[15] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 1, 8

[16] Netanel Kligler, Sagi Katz, and Ayellet Tal. Document enhancement using visibility detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2374–2382, 2018. 7, 8

[17] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 8

[18] Deng Li, Yue Wu, and Yicong Zhou. Sauvolanet: learning adaptive sauvola network for degraded document binarization. In *International Conference on Document Analysis and Recognition*, pages 538–553. Springer, 2021. 3

[19] Gen Li, Varun Jampani, Laura Sevilla-Lara, Deqing Sun, Jonghyun Kim, and Joongkyu Kim. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8334–8343, 2021. 8

[20] Yun-Hsuan Lin, Wen-Chin Chen, and Yung-Yu Chuang. Bedsr-net: A deep shadow removal network from a single document image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12905–12914, 2020. 3, 7, 8

[21] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13365–13374, 2021. 6

[22] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 8

[23] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 8

[24] 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*, pages 10684–10695, 2022. 2, 8

[25] 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. *arXiv preprint arXiv:2208.12242*, 2022. 8

[26] 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, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 8

[27] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion

probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 8

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 2

[29] Monika Sharma, Abhishek Verma, and Lovekesh Vig. Learning to clean: A gan perspective. In *Asian Conference on Computer Vision*, pages 174–185. Springer, 2018. 1, 3, 6, 8

[30] Daqian Shi, Xiaolei Diao, Lida Shi, Hao Tang, Yang Chi, Chuntao Li, and Hao Xu. Charformer: A glyph fusion based attentive framework for high-precision character image denoising. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1147–1155, 2022. 2, 3, 6, 8

[31] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3, 8

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 4, 8

[33] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 8

[34] 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*, 2021. 1, 2

[35] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 3, 4, 8

[36] Mohamed Ali Souibgui, Sanket Biswas, Sana Khamekhem Jemni, Yousri Kessentini, Alicia Fornés, Josep Lladós, and Umapada Pal. Docentr: an end-to-end document image enhancement transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1699–1705. IEEE, 2022. 3, 8

[37] Mohamed Ali Souibgui and Yousri Kessentini. De-gan: a conditional generative adversarial network for document enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3, 8

[38] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022. 1, 2

[39] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1788–1797, 2018. 7, 8

[40] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 6

[41] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 2, 8

[42] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[43] Xiangyu Xu, Deqing Sun, Jinshan Pan, Yujin Zhang, Hanspeter Pfister, and Ming-Hsuan Yang. Learning to super-resolve blurry face and text images. In *Proceedings of the IEEE international conference on computer vision*, pages 251–260, 2017. 3, 8

[44] Jiulong Zhang, Mingtao Guo, and Jianping Fan. A novel generative adversarial net for calligraphic tablet images denoising. *Multimedia Tools and Applications*, 79(1):119–140, 2020. 6

[45] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1

[46] Guoping Zhao, Jiajun Liu, Jiacheng Jiang, Hua Guan, and Ji-Rong Wen. Skip-connected deep convolutional autoencoder for restoration of document images. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2935–2940. IEEE, 2018. 3, 8

[47] Min Zhao, Fan Bao, Chongxuan Li, and Jun Zhu. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *arXiv preprint arXiv:2207.06635*, 2022. 8

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 1, 3, 6