João Gabriel Vinholi[1], Marco Chini[1], Anis Amziane[1], Renato Machado[1], Danilo Silva[1], and Patrick Matgen[1]

[1]Affiliation not available

April 16, 2024

Posted on 16 Apr 2024 — CC-BY-NC-SA 4 — https://doi.org/10.36227/techrxiv.171328129.97828901/v1 — e-Prints posted on TechRxiv are preliminary reports that are not peer reviewed. They should not b...

Optical Image-to-Image Translation Using Denoising Diffusion Models: Heterogeneous Change Detection as a Use Case

João Gabriel Vinholi, Marco Chini, *Senior Member, IEEE*, Anis Amziane, Renato Machado, *Senior Member, IEEE*, Danilo Silva, *Member, IEEE*, Patrick Matgen

***Abstract*—We introduce an innovative deep learning-based method that uses a denoising diffusion-based model to translate low-resolution images to high-resolution ones from different optical sensors while preserving the contents and avoiding undesired artifacts. The proposed method is trained and tested on a large and diverse data set of paired Sentinel-II and Planet Dove images. We show that it can solve serious image generation issues observed when the popular classifier-free guided Denoising Diffusion Implicit Model (DDIM) framework is used in the task of Image-to-Image Translation of multi-sensor optical remote sensing images and that it can generate large images with highly consistent patches, both in colors and in features. Moreover, we demonstrate how our method improves heterogeneous change detection results in two urban areas: Beirut, Lebanon, and Austin, USA. Our contributions are: i) a new training and testing algorithm based on denoising diffusion models for optical image translation; ii) a comprehensive image quality evaluation and ablation study; iii) a comparison with the classifier-free guided DDIM framework; and iv) change detection experiments on heterogeneous data.**

I. Introduction

IMAGE-TO-IMAGE translation (I2I) is a technique that transforms an image from one domain to another while preserving its original content [1]–[8]. This is useful for various applications such as super-resolution, inpainting, image decompression, and domain adaptation. In the context of remote sensing, I2I is useful when images acquired from different sensors need to be compared. For example, a deep learning model that learns from paired images can translate a low-resolution optical image to a high-resolution optical image or vice versa. Thus, synthetic versions of high-resolution images from low-resolution ones can be created, which are useful when the original high-resolution images are not available or too costly to acquire. However, I2I in this context is challenging because of the differences in spatial and spectral characteristics, noise levels, and geometric distortions of the heterogeneous images. Therefore, an I2I model needs to learn how to preserve the content and structure of the source image while adapting to the style and resolution of the target domain.

We address the shortcomings of existing I2I methods [3], [7], [9]–[11] for remote sensing imaging, which focus on patch-level translation and do not account for the variability of optical images due to multiple factors. Also, so far, no other work has proposed an algorithm that performs domain adaptation and super-resolution as a singular task using denoising diffusion models (DDM). A novel optical-to-optical translation approach is hereby propose. It leverages Denoising Diffusion Models (DDM) [12], which have superior image generation performance and avoid the problems of GANs and their variants [13], such as mode-collapse and undesired hallucinations. We aim to achieve domain-robust and high-quality I2I translation using DDMs, which are used by several leading AI companies for image generation.

The proposed DDM-based method differs from the standard framework commonly used in traditional computer vision tasks [1], [2], [14]–[17], as we have observed that the latter produces results that are highly inconsistent with neighboring patches, creating checkerboard-like images when patches are combined for a larger area. In contribution to this effect, the classifier-guided Denoising Diffusion Implicit Model (DDIM) [14], [15] seems to get confused by the high variability of colors between training patches when their features are very similar. This causes the generation of patches whose overall colorization is highly inaccurate. Our method tackle this problem by using color-standardization pre and post processing procedures carefully tailored to the optical-to-optical translation task.

Moreover, due to its stochastic nature, denoising diffusion models produce different outputs for a given input [12]. Even when not adding noise in between reverse diffusion steps [15], different choices of the initial random noise matrix result in different outputs for the same input. This characteristic, while extremely useful for tasks like inpainting and unconditional generation, can bring uncertainty about the consistency of generations in the context of I2I translation for remote sensing. In our experiments, we observed that the randomness brought by the initial noise matrix translates into the creation of a few but non-negligible low-quality generations caused by bad

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.
This work has been supported by the Luxembourg Directorate of Defence through the Chameleon project.
João Gabriel Vinholi, Marco Chini, Anis Amziane, and Patrick Matgen are with the Environmental Research and Innovation (ERIN) department of the Luxembourg Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg. E-mails: {joao.vinholi, marco.chini, anis.amziane, patrick.matgen}@list.lu
João Gabriel Vinholi is also with the Department of Telecommunications, Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, Brazil.
Renato Machado is with the Department of Telecommunications, Instituto Tecnológico de Aeronáutica (ITA), São José dos Campos, SP, Brazil. E-mail: rmachado@ita.br
Danilo Silva is with the Department of Electric and Electronic Engineering, Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brazil. E-mail: danilo.silva@ufsc.br

choices of initial noise conditions. These generations can be inconsistent with neighboring patches and with the features present in its correspondent input patch. To alleviate this, we propose a modification in the DDIM inference procedure that optimizes the choice of the initial random noise matrix.

As experiments, we apply our DDM-based method to generate synthetic high-resolution images for eight regions of interest: Beirut, Austin, Tlaquepaque, and five other locations. We evaluate the quality of the images produced by the proposed model using Learned Perceptual Image Patch Similarity (LPIPS) [18], Fréchet Inception Distance (FID) [19], and PSNR [20]. We find that our I2I method shows great improvement in image quality when using Sentinel-II as input, compared to the original Sentinel-II images and to synthetic images generated by deep regression-based models. To test the method in a practical application, we perform two heterogeneous change detection (HCD) experiments. HCD is a technique that compares images acquired from different sensors or different modes of image acquisition to identify changes in a specific area. Valuable for quick disaster damage assessment, it is, however, challenging to achieve accurate and consistent results due to the differences in image characteristics and quality [21]–[26]. The presented HCD tests show how our proposed I2I method can alleviate the effect of such differences. These experiments are performed on two regions: Beirut, Lebanon, and Austin, Texas, using a targetless change detection algorithm based on classical image processing tools. We compare the HCD results with and without the synthetic images. Our results demonstrate that the synthetic images generated by the proposed I2I method improve the HCD performance by greatly reducing false alarms and enhancing the change regions. Moreover, the synthetic images generated for this test are realistic, detailed, consistent, and with better contrast and clarity, compared to the original low-resolution images.

We summarize the main contributions of this work as:

1) The proposal of training and testing procedures based on denoising diffusion models that are able to translate optical images from a lower spatial resolution to a higher spatial resolution domain while maintaining high inter-patch and input-output consistency.
2) The presentation of image quality comparisons between the proposed method and other solutions, together with ablation experiments, over hundreds of square kilometers of imagery.
3) The execution of tests that show how one of the most ubiquitous frameworks for diffusion models, the classifier-free guided DDIM [14], might not be well suited to perform I2I translation of optical remote sensing images when trained with large and diverse data sets, and how our method is not affected by the same problems.
4) The presentation of heterogeneous change detection tests in Beirut, Lebanon, and Austin, USA, indicate that the proposed I2I translation method has the potential to improve heterogeneous change detection performance, compared to when no translation between domains is performed.

A. Related Works

1) Denoising Diffusion Models: Ho et al. [12] proposed the use of a denoising-based framework for image generation, the so-called Denoising Diffusion Probabilistic Model (DDPM). It showed incredible image generation quality and overcame most state-of-the-art methods available at the time. However, the inference process was highly inefficient since hundreds of model evaluations per image were necessary to reach optimal performance. The work *Denoising Diffusion Implicit Models* (DDIM) [15] approached this issue by presenting a new interpretation of the reverse diffusion process. Differently from the work in [12], they treated reverse diffusion as a non-Markovian process by skipping multiple reverse diffusion steps. Furthermore, they experimented with eliminating the randomness of intermediate diffusion steps—which came from the addition of a random noise matrix to each reverse diffusion step result—and noticed that this enabled an even greater reduction of model evaluations during inference at the cost of sample variety. In practice, these combined modifications reduced the necessary number of model evaluations during inference by up to two orders of magnitude while maintaining high-generation quality.

As originally defined in [12], denoising diffusion models are unconditional, i.e., they are trained to generate images with distributions similar to those it has seen during training without any prior knowledge about how a particular image should resemble or what it should contain. This limited the applicability of such models in multiple condition-guided tasks, such as image inpainting, super-resolution, colorization, and domain adaptation. To make them possible, two works [1], [16] experimented with adding a conditioning matrix as an additional model input. This partially worked since, as observed by Dhariwal et al. [2], diffusion models have a tendency to ignore the condition and focus only on the noisy input to be denoised. The work in [2] also proposed a possible strategy to alleviate this issue: the use of a separate classifier through which conditioning information is directly inserted into the reverse diffusion procedure. This proved to be useful, as the images generated with this procedure were strongly linked with the conditions added to the model. Still, the need to have a second classifier model is cumbersome and sometimes impractical. Due to that, Ho et al. [14] proposed *Classifier-free Guidance*, which, as the name implies, removes the need for a separate classifier network that guides the model with conditioning information. They accomplish that by making a small modification in the DDPM training and inference procedures that, in summary, treats the diffusion model as if it consisted of a combination of an unconditional and a conditional model. Hence, during inference, it became possible to determine how much the conditional and unconditional virtual parts of the model influenced the generation and, therefore, how strongly the reverse diffusion should rely on the conditioning information.

2) Image-to-Image Translation: In the context of image-to-image translation, Saharia et al. [1] investigated the use of DDPMs in different I2I translation tasks: colorization, image restoration, inpainting, and uncropping, obtaining re-

markable image generation quality and high sample diversity. Moreover, in the presented super-resolution tests, it beat deep regression-based models by an impressively high margin. The work in [17] used classifier-free guidance [14] in the task of semantic image generation—where the user inputs the information on what each area of the generated image should contain—together with a novel conditional model architecture that includes multi-head attention-based mechanisms. In the considered test sets, they beat all state-of-the-art methods in image quality metrics. *Image Super-Resolution via Iterative Refinement* [27] employed DDPM in the task of super-resolution. They trained a conditional model, referred to as SR3, to generate high-resolution samples starting from their low-resolution versions, beating all compared methods in image quality metrics and fool rate tests.

Other works focused on remote sensing-related tasks. Ismael et al. [5] proposed a GAN-based I2I translation method with the goal of performing remote sensing semantic segmentation of optical images from unknown domains using a segmentation network trained with optical images from a single domain. The methods in [6] and [8] proposed style-transferring domain adapting frameworks that generate extra training data matching the appearance of a particular same-resolution domain, to alleviate train/test data drift. The method in [10] uses a cycle-consistent GAN to translate optical images into SAR images, to perform change detection. It showed promising change detection results. However, training and testing data are limited to just a few acquisitions. The method in [7] is a thermodynamics-inspired translation network that obtained interesting results in SAR-to-Optical translation tasks with its unconventional design. The work in [3] proposes a DDM-based Brownian Bridge method for SAR-to-Optical I2I translation. There, the authors suggest the use of a modified forward diffusion procedure that, instead of resulting in a pure-noise image, the target SAR image is iteratively transformed into the source optical image. They reason that this could help the model more easily understand the relationships across domains. Xiao et al. propose a diffusion-based super-resolution method that seeks to further explore prior information from the low-resolution input. When tested on a small-scale data set, it performed well.

Despite the advancements in image-to-image translation within the context of remote sensing, existing literature predominantly focuses on the generation of additional training data for direct use in posterior deep learning algorithms [5], [6], [8], or in SAR to Optical/Optical to SAR translation [3], [7], [10]. These methods, while highly valuable, do not address the challenge of translating images across disparate optical sensors with different spatial resolutions, particularly ensuring feature consistency between input and output images. This gap is significant as it pertains to the fidelity of the translated images, where the preservation of salient features without introducing non-existent elements is crucial. Our work pioneers in this direction by proposing a novel framework that not only translates images from varying-resolution optical sensors but also guarantees the consistency of features. This approach mitigates the common issue of model hallucinations, where the generation of artifacts or the alteration of key elements could lead to misinterpretations, especially in critical applications like environmental monitoring and urban planning. By maintaining strict adherence to feature consistency, our method ensures that the output images faithfully represent the true characteristics of the input.

II. Proposed Method

Our method tackles the challenging task of optical-to-optical image translation between a lower spatial resolution sensor domain and a higher spatial resolution sensor domain. While higher spatial resolution is the most noticeable difference between such domains, there are also many other subtle features that might be difficult to pinpoint. The method hereby described leverages denoising probabilistic diffusion models to not only virtually increase spatial resolution but also to match these less obvious but important differences between the domains. This requires a training dataset with pairs of co-registered low spatial resolution (LR) domain A and high spatial resolution (HR) domain B images, showing the same scenes taken with different sensors, to ensure a thorough domain adaptation.

The full method generates multi-patch images with highly consistent and color-accurate patches. These images are sharp and clear, showing significant visual resolution improvements compared to lower-resolution domain images while preserving the original content. Our experiments in Section V provide evidence of these improvements.

Algorithm 1 Patchwise Training Step

Input: $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$, $\boldsymbol{I}_{\text{in}}^{\text{LR,global}}$, $\boldsymbol{I}_{\text{out}}^{\text{HR}}$, $\boldsymbol{\gamma}$, p_{uncond}, $\boldsymbol{\theta}_\tau$ {The model at training step τ}

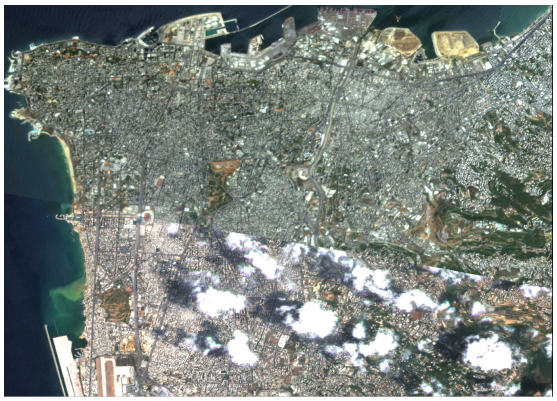
Output: $\boldsymbol{\theta}_{\tau+1}$

$\boldsymbol{I}_{\text{in}}^{\text{LR}} = \left[\boldsymbol{I}_{\text{in}}^{\text{LR,local}}; \boldsymbol{I}_{\text{in}}^{\text{LR,global}}\right]$
$u \sim \text{Bernoulli}(p_{\text{uncond}})$
if $u \equiv 1$ **then**
  $\boldsymbol{x} = \emptyset_{2n_{\text{ch}} \times h \times w}$
else
  $\{\boldsymbol{x}, _, _, _\} = \mathbf{W}\left(\boldsymbol{I}_{\text{in}}^{\text{LR}}, 2n_{\text{ch}}, h, w\right)$
end if
$\{\boldsymbol{y}_0, _, _, _\} = \mathbf{W}\left(\boldsymbol{I}_{\text{out}}^{\text{HR}}, n_{\text{ch}}, h, w\right)$
$\boldsymbol{\varepsilon} = (\varepsilon_{i,j,k})_{1 \le i \le n_{\text{ch}},\, 1 \le j \le h,\, 1 \le k \le w}$
  $\varepsilon_{i,j,k} \sim \mathcal{N}(0,1)\ \forall\, i,j,k$
$\gamma \sim \text{Uniform}(\boldsymbol{\gamma})$
$\tilde{\boldsymbol{y}} \leftarrow \sqrt{\gamma}\boldsymbol{y}_0 + \sqrt{1-\gamma}\boldsymbol{\varepsilon}$ {Forward diffusion}
$\hat{\boldsymbol{\varepsilon}} \leftarrow \boldsymbol{\theta}_\tau\left(\tilde{\boldsymbol{y}}, \boldsymbol{x}\right)$
$\boldsymbol{\theta}_{\tau+1} \leftarrow$ Optimization step on $\nabla_{\boldsymbol{\theta}_\tau} \mathcal{L}(\boldsymbol{\varepsilon}, \hat{\boldsymbol{\varepsilon}})$

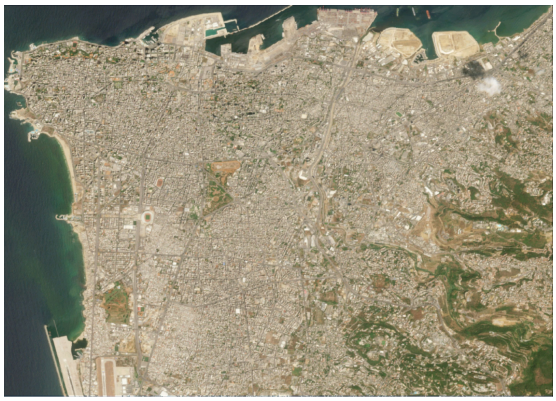
A. Training

Our training procedure diverges from conventional Denoising Diffusion Probabilistic Models (DDPM) by incorporating classifier-free guidance training [14]. This addition enforces

(a) Beirut Pre Setninel-II



(b) Beirut Pre Planet Dove



(c) Beirut Post Setninel-II

(d) Beirut Post Planet Dove

Fig. 1: Satellite images from Beirut, Lebanon, of low and high resolution, captured at a very similar time interval, from Sentinel-II and Planet Dove sensors. Figures 1a and 1c are the pre and post-event images from Sentinel-II, respectively. Figures 1b and 1d are the pre and post-event images from Planet Dove, respectively.

Algorithm 2 Image Whitening $\mathbf{W}$

Input: $\boldsymbol{I}_{\mathrm{c}}$, n_{ch}, h, w

Output: $\boldsymbol{I}_{\mathrm{w}}$, $\boldsymbol{m}_1, m_2, m_3$

$\boldsymbol{I}_{\mathrm{w}} \leftarrow \mathbf{0}_{n_{\mathrm{ch}} \times w \times h}$

$\boldsymbol{m}_1 \leftarrow \mathbf{0}_{n_{\mathrm{ch}}}$

$m_2 \leftarrow 0$

$m_3 \leftarrow 0$

for $i \leftarrow 0$ to $n_{\mathrm{ch}} - 1$ **do**

  $m_1^i \leftarrow \mu(\boldsymbol{I}_{\mathrm{c}}^i)$

  $\boldsymbol{I}_{\mathrm{w}}^i \leftarrow \boldsymbol{I}_{\mathrm{c}}^i - m_1^i \cdot \mathbf{J}_{h \times w}$ $\{\mathbf{J}_{h \times w}$ is the all-ones $h \times w$ matrix$\}$

end for

$m_2 \leftarrow \min \boldsymbol{I}_{\mathrm{w}}$

$\boldsymbol{I}_{\mathrm{w}} \leftarrow \boldsymbol{I}_{\mathrm{w}} - m_2 \cdot \mathbf{J}_{n_{\mathrm{ch}} \times h \times w}$

$m_3 \leftarrow \max \boldsymbol{I}_{\mathrm{w}}$

$\boldsymbol{I}_{\mathrm{w}} \leftarrow 2\,(\boldsymbol{I}_{\mathrm{w}}/m_3 - 0.5 \cdot \mathbf{J}_{n_{\mathrm{ch}} \times h \times w})$

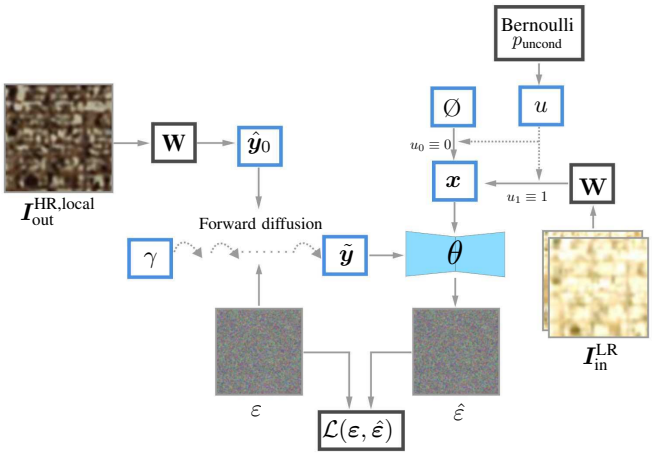


Fig. 2: Diagram illustrating the training step procedure for the proposed DDM-based image-to-image translation method.

alignment between input and output patches, thereby mitigating the risk of model hallucinations. Additionally, we employ color standardization, a pre-and a post-processing technique: Whitening and Coloring, respectively. These steps are critical as they prevent the diffusion network from being confounded

by the high variability in patch coloration found within extensive training datasets, particularly when the patches have highly similar content. Without these measures, the network struggles to accurately predict the channel mean values for the output patches, as shown in Section V.

Algorithm 1 defines the training step for a single pair of low spatial resolution patch from domain A, $\boldsymbol{I}_{\text{in}}^{\text{LR}} \in \mathbb{R}^{2n_{\text{ch}} \times h \times w}$, and high resolution patch, $\boldsymbol{I}_{\text{out}}^{\text{HR}} \in \mathbb{R}^{n_{\text{ch}} \times h \times w}$, from domain B. The goal is to train a model $\boldsymbol{\theta}$ that is able to convert $\boldsymbol{I}_{\text{in}}^{\text{LR, local}}$ into $\boldsymbol{I}_{\text{out}}^{\text{HR}}$. The training step is illustrated by Figure 2. The input and output images have the same pixel resolution but differ in spatial resolution, i.e., $\boldsymbol{I}_{\text{in}}^{\text{LR, local}}$ is up-scaled with cubic interpolation to have the pixel resolution of $\boldsymbol{I}_{\text{out}}^{\text{HR}}$, while covering the exact same area and the same features. Thus, patches aligned in time are required for training to work. Line 1 of the algorithm defines $\boldsymbol{I}_{\text{in}}^{\text{LR}}$ as the channel-wise stacking of $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$ and $\boldsymbol{I}_{\text{in}}^{\text{LR, global}}$, which are, respectively, the region of interest (ROI) of the low-resolution image from domain A, which is the patch to be translated to domain B in inference, and the surroundings of the region of interest, down-sampled to have the same dimensions of $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$. $\boldsymbol{I}_{\text{in}}^{\text{LR, global}}$ is defined to contain four times the visible area of $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$, i.e., the area that is contained in $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$ plus three additional equally sized regions in its vicinity, thus depicting north, east and southeast regions that are connected to $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$. This is done to give more context to the model of what exactly lies around the input patch. Line 2 samples a random variable u from a Bernoulli distribution of probability p_{uncond}. If it turns out to be 1, the model behaves as an unconditional diffusion model for the current training step (Line 4), as in [14]. Otherwise, it behaves as a conditional diffusion model. $\emptyset$ is the null-condition matrix, as also defined in [14]. To ensure that $\emptyset$ is correctly interpreted by the model, we set it as a matrix whose all elements are equal to -2, which is outside the range of values of input pixels $[-1, 1]$. Thus, the model does not interpret the null condition as a regular input image and vice-versa. Lines 6 and 8 define the conditioning input from domain A, $\boldsymbol{x} \in \mathbb{R}^{2n_{\text{ch}} \times h \times w}$, and the domain B ground truth $\boldsymbol{y}_0 \in \mathbb{R}^{n_{\text{ch}} \times h \times w}$, respectively. They are obtained by applying the function $\mathbf{W}$, defined in Algorithm 2, to $\boldsymbol{I}_{\text{in}}^{\text{LR}}$ and $\boldsymbol{I}_{\text{out}}^{\text{HR}}$. This function extracts out the overall color information from an input image, i.e., it zeroes the mean of each channel. A normalization between $[-1, 1]$ follows. $\mathbf{W}$ also returns the previous means, minimum, and maximum values extracted from the input, which can be used for reverting the whitening process during inference. However, during training, these values are not used. Lines 9 to 13 of Algorithm 1 are as in the original conditional DDPM training procedure [1]. This process is repeated for all existing training patches until the learning has converged.

B. Inference

The inference procedure, illustrated by Figure 3, aims at maximizing the image generation quality by selecting the best possible initial noise matrix to be fed to the reverse diffusion process. We adapt the classifier-free guided DDIM approach by integrating the same color standardization methods used

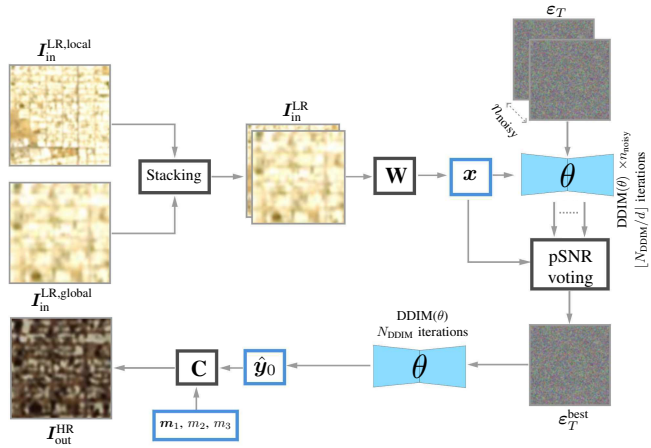


Fig. 3: Diagram illustrating the inference procedure for the proposed DDM-based image-to-image translation method. Here, the color information variables $\boldsymbol{m}_1$, m_2, m_3 are provided from an external image, e.g., a post-event Planet Dove image.

Algorithm 3 Patchwise Inference

Input: $\boldsymbol{\theta}, \boldsymbol{I}_{\text{in}}^{\text{LR,local}}$, $\boldsymbol{I}_{\text{in}}^{\text{LR,global}}$, n_{ch}, h, w, ω_{uncond}, N_{DDIM}, d, n_{noisy}

Optional: $\boldsymbol{m}_1$, m_2, m_3 {Color information variables.}

Output: $\boldsymbol{I}_{\text{out}}^{\text{HR,local}}$

1: $\boldsymbol{I}_{\text{in}}^{\text{LR}} = \left[\boldsymbol{I}_{\text{in}}^{\text{LR,local}}\ \boldsymbol{I}_{\text{in}}^{\text{LR,global}}\right]^T$
2: **if** $\boldsymbol{m}_1$, m_2, m_3 are all provided **then**
3: $\quad \{\boldsymbol{x}_{\text{local}}, _, _, _\} = \mathbf{W}\left(\boldsymbol{I}_{\text{in}}^{\text{LR,local}}, n_{\text{ch}}, w, h\right)$
4: $\quad \{\boldsymbol{x}_{\text{global}}, _, _, _\} = \mathbf{W}\left(\boldsymbol{I}_{\text{in}}^{\text{LR,global}}, n_{\text{ch}}, w, h\right)$
5: **else**
6: $\quad \{\boldsymbol{x}_{\text{local}}, \boldsymbol{m}_1, m_2, m_3\} = \mathbf{W}\left(\boldsymbol{I}_{\text{in}}^{\text{LR,local}}, n_{\text{ch}}, w, h\right)$
7: $\quad \{\boldsymbol{x}_{\text{global}}, _, _, _\} = \mathbf{W}\left(\boldsymbol{I}_{\text{in}}^{\text{LR,global}}, n_{\text{ch}}, w, h\right)$
8: **end if**
9: $\boldsymbol{x} = [\boldsymbol{x}_{\text{local}};\ \boldsymbol{x}_{\text{global}}]$
10: $\boldsymbol{\mathcal{E}}_T \leftarrow \left(\varepsilon_T^{i,j,k,l}\right)_{1 \le i \le n_{\text{noisy}},\ 1 \le j \le n_{\text{ch}},\ 1 \le k \le h,\ 1 \le l \le w}$
$\quad \varepsilon_T^{i,j,k,l} \sim \mathcal{N}(0,1)\ \forall\, i, j, k, l$
11: $\hat{\boldsymbol{Y}}_{0,\text{lowqual}} \leftarrow \mathbf{0}_{n_{\text{noisy}} \times n_{\text{ch}} \times h \times w}$
12: **for** $i \leftarrow 1$ to n_{noisy} **do**
13: $\quad \hat{\boldsymbol{Y}}_{0,\text{lowqual}}^{i} \leftarrow \text{DDIM}\left(\boldsymbol{\theta}, \boldsymbol{\mathcal{E}}_T^{i}, \omega_{\text{uncond}}, \boldsymbol{x}, \lfloor N_{\text{DDIM}}/d \rfloor\right)$
14: **end for**
15: $\boldsymbol{\mathcal{E}}_T^{\text{best}} = \arg\max_{\boldsymbol{\mathcal{E}}_T^{j}} \text{PSNR}\left(\boldsymbol{x}_{\text{local}}, \hat{\boldsymbol{Y}}_{0,\text{lowqual}}^{i}\right)$
16: $\hat{\boldsymbol{y}}_0 \leftarrow \text{DDIM}\left(\boldsymbol{\theta}, \boldsymbol{\mathcal{E}}_T^{\text{best}}, \omega_{\text{uncond}}, \boldsymbol{x}, N_{\text{DDIM}}\right)$
17: $\boldsymbol{I}_{\text{out}}^{\text{HR,local}} = \mathbf{C}(\hat{\boldsymbol{y}}_0, n_{\text{ch}}, h, w, \boldsymbol{m}_1, m_2, m_3)$

during training. To further enhance the fidelity of the generated patches to their inputs, we refine the selection process for the initial noise condition input for the reverse diffusion process, ensuring optimal consistency and translation quality.

In our experiments, we observed that, for the same conditioning input $\boldsymbol{x}$, the quality of DDIM reverse diffusion outputs generated with different initial noise matrices varied significantly. That is, for a combination of independent and identically distributed random initial noise matrices, fixing $\boldsymbol{x}$, the DDIM, in some cases, generates samples entirely consistent with $\boldsymbol{x}$, whereas sometimes it generates inconsistent samples with undesired hallucinations. Since the DDIM reverse diffusion is a deterministic process, i.e., given fixed $\boldsymbol{\varepsilon}_T$ and $\boldsymbol{x}$, its output remains the same, the cause for varying consistency levels lies only in the different sampled initial noise matrices, which are the source of the randomness of the whole inference process. Thus, we propose a procedure that seeks to select an initial noise matrix such that the DDIM output possesses the highest consistency compared to $\boldsymbol{x}$. This procedure consists of repeating the DDIM reverse diffusion process with a reduced number of iterations, using as input n_{noisy} different initial noise matrices. The resulting outputs are tested for consistency by calculating the peak signal-to-noise ratio (PSNR) between $\boldsymbol{x}_{\text{local}}$ and each noise matrix individually. The one that was responsible for the maximum PSNR is used as the definitive starting noise matrix for the inference process, which takes place with a higher number of iterations, to ensure maximal generation quality.

Algorithm 3 defines the inference process for a single low spatial resolution domain A patch $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$, which results in a translated high spatial resolution patch $\boldsymbol{I}_{\text{out}}^{\text{HR}}$ resembling those of domain B, containing the same elements of $\boldsymbol{I}_{\text{in}}^{\text{LR, local}}$. As in Algorithm 1, the input and output images have the same pixel resolution but differ in spatial resolution, meaning that the lower-resolution input is up-sampled to the output's pixel dimensions. In the first line of the algorithm, we define the conditioning input $\boldsymbol{I}_{\text{in}}^{\text{LR}}$ as the concatenation of the domain A patch to be translated $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$ and its down-sampled surroundings, as done in training. Lines 1 to 8 prepare the low spatial resolution domain A input to be in the format required by the trained neural network $\boldsymbol{\theta}$. The color information-related variables $\boldsymbol{m}_1, m_2, m_3$ are assigned, which are extracted during the whitening process of the local domain A patch if the colors from the input patch are desired to be reassigned to the output image from domain B. Otherwise, color information from an external image, e.g., a post-event domain B image, can be used. Following, line 10 defines $\boldsymbol{\mathcal{E}}_T$ as a Gaussian noise matrix that can be interpreted as n_{noisy} individual matrices stacked together in a new dimension. Each of these is used as initial noise matrices for the DDIM reverse diffusion process [15] described in line 13. Thus, DDIM inference is executed n_{noisy} times using the trained model $\boldsymbol{\theta}$, with different initial noise matrices $\boldsymbol{\mathcal{E}}_T^i$, but with a small number of iterations $\lfloor N_{\text{DDIM}}/d \rfloor$. It performs classifier-free guidance inference using ω_{uncond} as the weighting parameter, which controls how much the unconditional score estimate is weighted relative to the conditional score estimate. A smaller ω_{uncond} means more weight of the unconditional score estimate, which can increase the mode coverage and diversity of the generated images, but it can also decrease the sample fidelity and quality. Each of the n_{noisy} inferences generates a prediction $\hat{\boldsymbol{Y}}_{0,\text{lowqual}}^i$, that, due to the lower number of DDIM iterations, are coarse predictions, whose image quality is nevertheless good enough for the following step: they are used to evaluate which of the n_{noisy} noise matrices contained in $\boldsymbol{\mathcal{E}}_T$ created the prediction of highest quality with respect to the peak signal-to-noise-ratio (PSNR) metric, comparing it to the first n_{ch} channels of the conditioning input $\boldsymbol{x}$, which refer to $\boldsymbol{I}_{\text{in}}^{\text{LR,local}}$. The chosen noise matrix is denoted as $\boldsymbol{\mathcal{E}}_T^{\text{best}}$, and is used to generate the final predicted output of the DDIM reverse diffusion $\hat{\boldsymbol{y}}_0$, now using N_{DDIM} iterations. Finally, the high spatial resolution output patch $\boldsymbol{I}_{\text{out}}^{\text{HR,local}}$ is obtained by applying Algorithm 4 to $\hat{\boldsymbol{y}}_0$, which gives the overall color information to the DDIM's prediction.

Algorithm 4 Image Colorization $\mathbf{C}$

Input: $\boldsymbol{I}_{\text{w}}$, n_{ch}, h, w, $\boldsymbol{m}_1$, m_2, m_3

Output: $\boldsymbol{I}_{\text{c}}$

1: $\boldsymbol{I}_{\text{c}} \leftarrow \mathbf{0}_{n_{\text{ch}} \times h \times w}$
2: $\boldsymbol{I}_{\text{w}} \leftarrow \frac{\boldsymbol{I}_{\text{w}} - \min \boldsymbol{I}_{\text{w}} \mathbf{J}_{n_{\text{ch}} \times h \times w}}{\max \boldsymbol{I}_{\text{w}} - \min \boldsymbol{I}_{\text{w}}}$
3: $\boldsymbol{I}_{\text{w}} \leftarrow m_3 \cdot \boldsymbol{I}_{\text{w}}$
4: $\boldsymbol{I}_{\text{w}} \leftarrow \boldsymbol{I}_{\text{w}} + m_2 \cdot \mathbf{J}_{n_{\text{ch}} \times h \times w}$
5: **for** $i \leftarrow 0$ to $n_{\text{ch}} - 1$ **do**
6: $\quad \boldsymbol{I}_{\text{c}}^i \leftarrow \boldsymbol{I}_{\text{w}}^i + m_1^i \cdot \mathbf{J}_{h \times w}$
7: **end for**

III. Model Setup

The model chosen for the experiments is the one presented in [17], where the authors propose a U-Net-style model that uses attention mechanisms to improve image fidelity. Differently from the diffusion-based models presented in [1], [14], [27], this model processes the noisy input $\tilde{\boldsymbol{y}}$ and the conditioning input $\boldsymbol{x}$ independently. That is, $\tilde{\boldsymbol{y}}$ is fed right into the encoder and is processed normally, whereas $\boldsymbol{x}$ is re-introduced to the decoder by inserting it into each residual block. To properly match the spatial resolution required by each residual block, $\boldsymbol{x}$ is processed by spatially adaptive normalization (SPADE) blocks. As stated by the authors of the model, we observed that this modification improved significantly the generation quality and fidelity to the conditioning input. Some internal parameters of the model were adjusted to maximize performance while limiting its size due to hardware constraints. The number of channels has been set to be a multiple of 96. The number of residual blocks for each downsampling operation has been set to 2. Attention resolutions were set to $2, 4, 8$ and 16. Finally, the number of attention heads has been set to 4. For training, batch size has been set to 5, and the learning rate has been set to 10^{-4}. We use the Huber loss function [29] with its parameter δ_ℓ set to 0.5 since it has been shown to reduce the impact of outliers in the gradients while maintaining the benefits of the mean square distance for non-outliers. RAdam optimizer [30] has been chosen due to its proven training speed and its low sensitivity to the learning rate choice. The training

TABLE I: Hyperparameter values and their purpose for the performed experiments.

Parameter	Value	Purpose
Patch Size	128×128	Width w and height h of the image patches fed to the neural network.
N_{DDIM}	64	Number of DDIM iterations for the final reverse diffusion in Line 16 of Alg. 3.
d	8	Constant used to reduce the number of DDIM iterations for the $\boldsymbol{\varepsilon}_T^{\text{best}}$ estimation in Line 15 of Alg. 3.
$\boldsymbol{\gamma}$	$\left[\cos\left(\frac{t_i/T+0.008}{1.008}\frac{\pi}{2}\right)^2\right]_{t_i}$	Diffusion noise level parameter for each timestep $t_i \in \{0, 1, \ldots, T-1\}$ [16].
T	1024	Number of forward diffusion noising steps.
n_{noisy}	8	Number of DDIM runs for the estimation of $\boldsymbol{\varepsilon}_T^{\text{best}}$, used in Line 13 of Alg. 3.
p_{uncond}	10^{-1}	Classifier-free guidance unconditional diffusion probability [14].
ω_{uncond}	1	Classifier-free guidance inference weighting parameter [14].
λ_{consist}	10^{-1}	Weight of the consistency loss over the final loss [28].
$\varepsilon_{\text{consist}}$	10	It tells how far from the sampled timestep t can the consistency timestep t' be sampled [28]. That is, $t' \in \{t - \varepsilon_{\text{consist}}, \ldots, t + \varepsilon_{\text{consist}}\}$.
n_{consist}	1	Number of consistency training steps to be executed in a single overall training step [28].
γ_{SNR}	5	Clamping parameter for the Min-SNR-γ training weighting strategy.

ran for 31 epochs, which was more than enough for the loss function to converge completely.

To stabilize training and avoid catastrophic forgetting of features, Exponential Moving Average (EMA) has been used in several DDM-based methods [1], [2], [12], [17]. With the same goal, we instead use Stochastic Weight Averaging (SWA) [31], which consists of averaging weights after the training loss shows convergence behavior. Differently from EMA, SWA does not give higher importance to the weights of more recent epochs, but instead considers all epochs after the beginning of weight averaging to have the same importance for the overall learning. Since we have observed that the loss function converged after a number of epochs, it was reasonable to consider that newer weight updates would not bring an overall improvement of the model, thus applying SWA instead of EMA made more sense. The work presented in [32] suggests that averaging weights of different epochs can indeed result in higher performance than when only using the last weights of a training experiment. We observed that training stabilized after 10 epochs, so we started the SWA process at this moment. Also, SWA suggests increasing the learning rate during the epochs where weight averaging takes place. However, we noticed that increasing the learning rate over 10^{-4} made training unstable. For this reason, it is kept constant during the whole process.

For the performed experiments, multiple hyperparameters not directly related to the model setup have been chosen. They are presented in Table I, along with their purpose. Some choices were based on a grid search, and others were based on the articles that defined them in the first place.

Two recently proposed training techniques for denoising diffusion models were included in our experiments: Consistency Diffusion Loss and Min-SNR-γ Weighting.

a) Consistency Diffusion Loss: It is a method to mitigate sampling drift in diffusion models by learning to be consistent [28]. It is based on enforcing a consistency property, which states that predictions of the model must be consistent across time when the conditioning input is the same. We observed that this significantly improved the consistency between neighbor patches of the same image frame and that it helped to reduce the frequency of undesired model hallucinations. The parameters related to consistency loss are:

- λ_{consist} This is the weight of the consistency loss over the final loss. It controls how much the consistency loss is used for learning relative to the DDPM loss.
- n_{consist}: This is the number of consistency training steps to be executed in a single overall training step.

b) Min-SNR-γ Weighting: It is a method to improve the convergence speed and effectiveness of DDPM training by adapting the loss weights of different timesteps of the diffusion process according to their signal-to-noise ratios (SNRs) [33]. It is based on the observation that the diffusion training suffers from conflicting optimization directions between timesteps, which slows down the learning process, as also visualized in our experiments. The single parameter of this method, γ, which hereby is referred to as γ_{SNR}, controls the degree of clamping the SNRs, which maintains stability and speed of the training process.

IV. Dataset

For the proposed experiments, paired images from satellites Sentinel-II and Planet Dove were gathered. The former is a widely known open-source optical satellite with a spatial resolution of 10 meters, whereas the latter is a class of commercial satellites aimed at producing high-quality images at a spatial resolution of 3 meters. A dataset containing 77

(a) Sentinel-II (b) Regression (c) Regression + $\mathbf{W}, \mathbf{C}$

(d) DDIM (e) DDIM + $\mathbf{W}, \mathbf{C}$ (f) DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting

Fig. 4: Comparison among images generated by the tested I2I methods (b)-(f) and the original Sentinel-II image (a). It is evident how the image generated with the proposed model (f) displays higher resolution features not observable among the others, while maintaining patch-wise feature consistency.

pairs of rasters has been assembled for training. These are parts of or a combination of acquisitions that were captured at similar time periods. They are properly geolocated and checked for image quality issues. In total, a vast area of 7894 km^2 is covered by the training set, which is comparable to the area of Puerto Rico (9104 km^2). As the test set for the image-to-image translation algorithm, 8 pairs of images were selected, which, in total, cover an area of 707 km^2. With these, image quality metrics are extracted for the proposed method and compared to other possible approaches for the I2I translation task.

For evaluation of the performance of the method in a change detection task, we use pre- and post-images from the port region in Beirut, Lebanon. On August 4th, 2020, an explosion dramatically affected the port and surrounding areas. As they have recovered over time, many changes took place, e.g., new and repaired constructions and structures. Thus, for the change detection test, we use as pre-event images a pair of images acquired at a time close to the explosion, between 04 and 20 of August 2020, when the effects of the event are still evident. As post images, acquisitions from July 2023 of the exact same region are used. Figure 1 shows the Beirut pre and post-event images from Sentinel-II and Planet Dove. We also perform change detection tests in a region in Austin, USA, where many urban changes can be observed between the selected pre- and post-event images. They were captured in July 2021 and 2023, respectively.

V. Experiments

A. Image Generation

1) Image Quality Metrics and Ablation Experiments: To effectively demonstrate the performance and advantages of the proposed I2I algorithm, a comparison in terms of image quality metrics is proposed. Table II displays the values of three image quality/comparison metrics: Frechét Inception Distance (FID), mean Peak Signal-to-Noise Ratio (PSNR), and mean Learned Perceptual Image Patch Similarity (LPIPS). To obtain these values, we use the test set described in Section IV. The FID [19] measures the similarity between two data sets of images—usually a data set of generated images and one of reference images—by computing the distance between their feature representations extracted by a pre-trained Inception network. A lower FID indicates better feature alignment of the generated images with the reference data set. The PSNR [20] measures the reconstruction error between two images in terms of the ratio of the maximum possible pixel value and the mean squared error. A higher PSNR indicates a lower reconstruction error and a better visual quality. It is important to note that

TABLE II: Image Quality Metrics of the Proposed Algorithm and Compared Approaches

Image Set	**mLPIPS↓**	**FID↓**	**mPSNR↑**
Sentinel-II	0.2977	98.08	12.76
Regression	0.2797	64.12	16.72
Regression + $\mathbf{W}, \mathbf{C}$	0.2491	67.04	17.48
Method in [14] (Conditional DDIM)	0.4769	87.93	9.919
Conditional DDIM + $\mathbf{W}, \mathbf{C}$	0.1993	46.89	15.39
Conditional DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting (Proposed)	0.1884	45.64	15.72

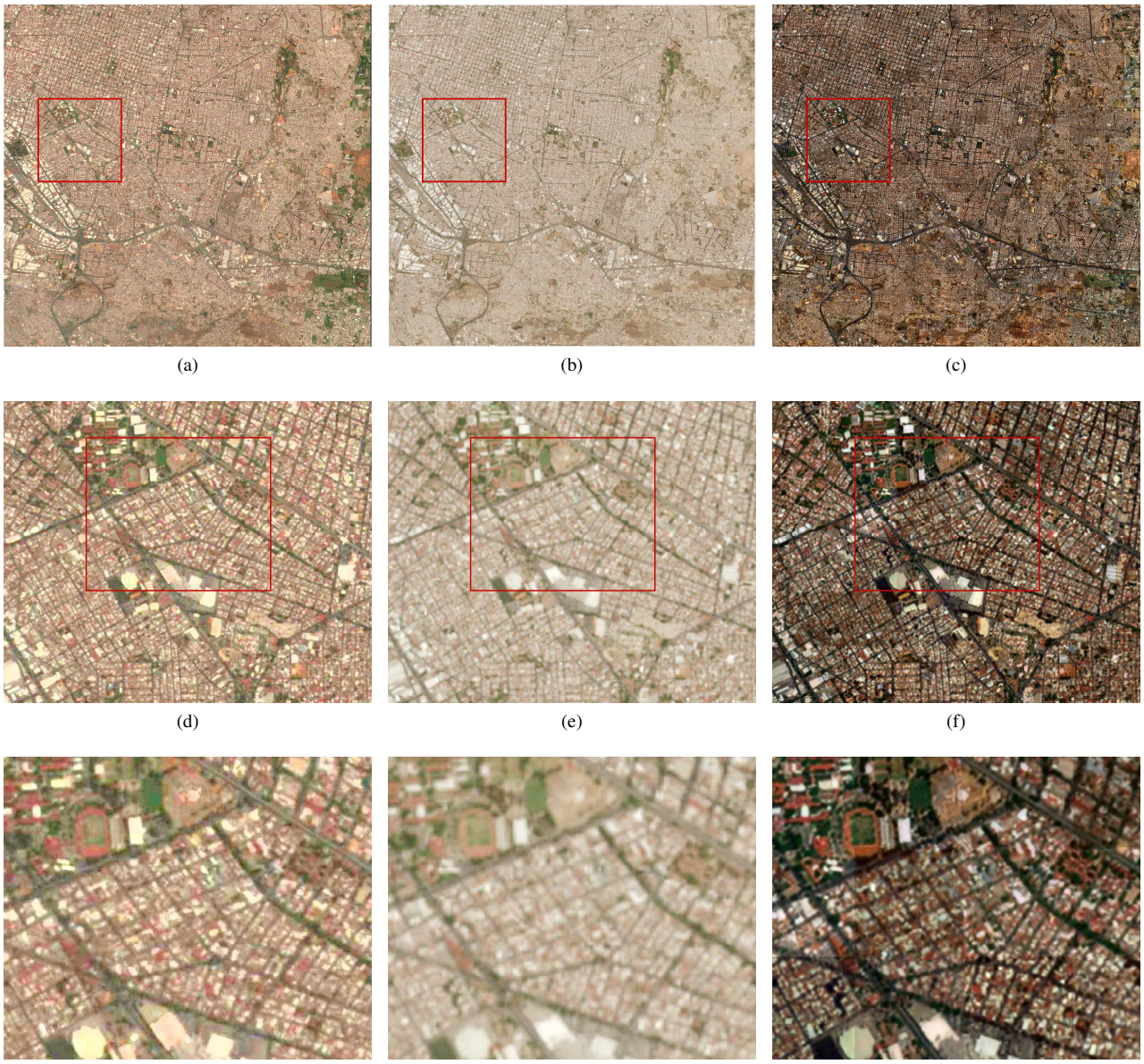


Fig. 5: Figures of the region of interest in Tlaquepaque, Mexico. In the first row, images of a far view of the city are displayed. In the second row, we can observe a cropped part of the images in the first row for closer inspection. At the third row, a cropped part of the images from the second row is displayed, for the inspection of meter-scale features. The first, second, and third columns contain images from Planet Dove, Sentinel-II, and Synthetic Planet domains, respectively.

PSNR is not very reliable for perceptual quality, as it does not account for the human visual system and the semantic content of the images [1]. Nonetheless, in the presented results, we compute the mean PSNR (mPSNR) of all Planet and Synthetic Planet corresponding patches, to further enrich the discussion. The LPIPS [18] measures the perceptual similarity between a pair of images by computing the distance between their feature representations extracted by a pre-trained deep neural network. A lower LPIPS indicates higher perceptual similarity and better visual quality. LPIPS is aligned with human perception, as it captures the semantic and structural differences between images [34]. As with the PSNR, we compute the mean LPIPS (mLPIPS) of all corresponding Planet and Synthetic Planet patches. In a patch-wise comparison, the mLPIPS metric is better suited as a comparison metric than FID since its calculation is done on a patch-by-patch basis instead of focusing on the whole test set at once. Thus, since our tests involve direct comparison between corresponding images, mLPIPS is also employed.

We compare our proposed I2I algorithm with other possible I2I solutions. First, the Sentinel-II test set is directly compared to Planet Dove images without any I2I translation method. The obtained metrics are shown in the first row of Table II. Second, we train two deep regression models [35] whose loss

objective is to directly match the pixel values of Planet patches when Sentinel-II corresponding patches are input. They share the same neural network backbone as the proposed DDM-based method. In other words, these models are identical to the one used in the proposed method, with the exception that the output linear layer directly predicts the values of pixels of the corresponding Planet patch for a given input. The first deep regression model, whose metrics are displayed in the second row of Table II, is a conventional regression model, i.e., it does not apply any pre or post-processing to its inputs and outputs apart from simple $[-1, 1]$ normalization. The second regression model, whose results are displayed in the third row, is identical to the previous model, except that the whitening and coloring algorithms, defined in Algorithm 2 and 4, are applied to its inputs and outputs, respectively. In the table, we refer to this as Regression + $\mathbf{W}, \mathbf{C}$. The goal of testing this configuration is to see what would be the influence of these processing methods in a conventional regression model and whether or not it has a significant impact on the metrics. Following, we test the standard classifier-free guided DDIM [14], i.e., without the proposed whitening, coloring, and PSNR voting processes, to perform an ablation of these techniques. It is referred to in Table II as Conditional DDIM. Then, we train a similar model, but now adding the whitening and coloring procedures. To measure the impact of the PSNR voting in the inference results, the image quality metrics for this model are separated into two rows, the first being without PSNR voting—denoted as Conditional DDIM + $\mathbf{W}, \mathbf{C}$—and the second with it—Conditional DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting.

Looking at the numbers of Table II, we observe that the full proposed technique, Conditional DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting, reaches, by a good margin, better results in mLPIPS and FID metrics when compared to the other methods. For instance, FID dropped by 52.44 points when compared to Sentinel-II images—an expressive improvement of 53.46%. When compared to Regression and Regression + $\mathbf{W}$, the FID has been respectively reduced by 18.48 and by 21.8 points—an improvement of 28.82% and 31.92%. Another interesting observation is that the vanilla DDIM could not beat the regression models in any of the metrics. Even worse, it loses to the original images in both mLPIPS and mPSNR. The empirical reason for this is that the vanilla diffusion training seems to get confused by the multiple different overall tonalities of patches containing similar features, which impacts the convergence ability of the model.

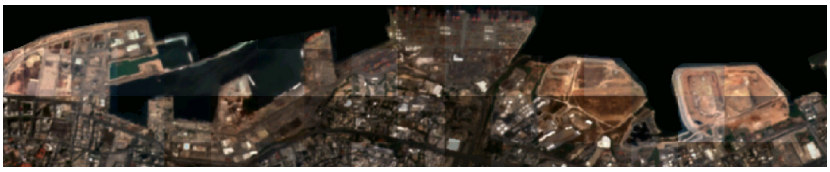
This confusion seems to be eliminated when whitening and coloring procedures are used, as evidenced by the performance superiority of both Conditional DDIM + $\mathbf{W}, \mathbf{C}$ and Conditional DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting. Looking at mLPIPS results, we observe that the regression-based models beat by a small margin the original Sentinel-II, whereas Conditional DDIM + $\mathbf{W}, \mathbf{C}$ and Conditional DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting show a much more significant improvement in the same regard. This indicates that, in a patch-by-patch basis, only DDIM + $\mathbf{W}, \mathbf{C}$ and DDIM + $\mathbf{W}, \mathbf{C}$ + PSNR Voting models were able to heavily reduce the perceptual distance of patches when compared to their Planet Dove versions. The proposed method with PSNR Voting has reduced mLPIPS by 0.1093 points, meaning that the produced patches are, on average, 36.71% more perceptually similar to their corresponding Planet Dove patches than their Sentinel-II versions. When we compare the last two rows of the table, PSNR voting brings a significant improvement in all metrics: -5.469% for mLPIPS, -2.665% for FID, and $+2.144\%$ for mPSNR. We observed, moreover, that PSNR voting is specially useful to discard rare but existent undesired model hallucinations and poor translations, which would negatively impact, for example, change detection results. These improvements, due to the increased number of model evaluations, come at the cost of higher numerical complexity.

While regression models achieve higher mPSNR values, as discussed, PSNR's preference for blurry regression-produced outputs doesn't align with human perception. Consequently, learning-based metrics like FID are the standard in image generation comparisons, rendering the PSNR performance superiority of regression-based models irrelevant.

2) Visual Examples: To properly convey the visual differences between the tested methods in Table II, Figure 4 shows generated examples of a region in Beirut, between (4b to 4f), and put them beside the corresponding Sentinel-II image (4a). As expected, regression-generated images (4b and 4c) have a blurry appearance and, due to this, when compared to the Sentinel-II version, an improvement in the apparent resolution is not easily noticeable. Figure 4b seems to have better inter-patch consistency, whereas the addition of whitening and coloring processes apparently brought some contrast enhancement in Figure 4c. Figure 4d is the result for the vanilla DDIM, where it can be noticed how the model has serious issues trying to figure out which is the overall tonality of the produced patches, thus creating a mosaic-like appearance when they are joined together to form a bigger image. This issue is solved by the inclusion of whitening and coloring processes, as observed by the result in Figure 4e. In fact, with their inclusion, 4e shows an even greater patch consistency than the regression-based results 4b and 4c. When compared to 4f, however, some regions of Figure 4e show deficiencies. In the center-west part of 4e, we observe a blurry and watched-out patch that apparently was not properly generated by the model. Likewise, in the south-west and south-east we observe two patches that are darker than their neighbors, and, for the south-east patch in particular, the darker patch also shows blurry low-resolution features that are not present in the same region in 4f. Overall, the PSNR voting procedure has improved patch consistency and suppressed non-ideal DDIM generations.

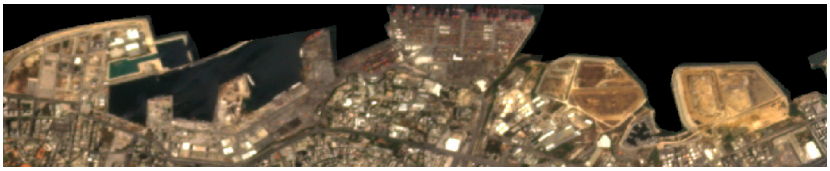
In a direct comparison between Sentinel-II, Planet Dove, and the synthetic Planet Dove image generated by the proposed method, Figure 5 presents a region from the city of Tlaquepaque, Mexico. Each of the three rows of figures shows a different view of the city, with different scales, to better compare different aspects of the images. The three columns show images from Planet Dove, Sentinel-II, and images generated by the proposed method, respectively. The first row shows a far view of the city that depicts multiple complex urban features: houses, buildings, storage facilities, factories, roads, and scattered vegetation. When looking at

(a) Pre-Event Sentinel-II

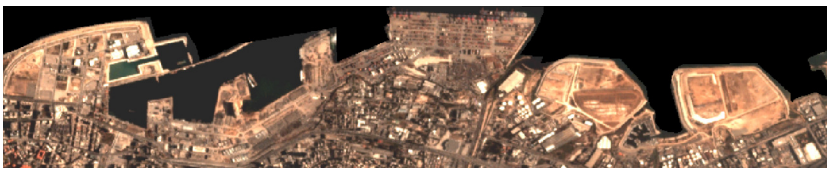


(b) Pre-Event Synthetic Planet Dove (Regression + $\mathbf{W}$)

(c) Pre-Event Synthetic Planet Dove (Proposed)



(d) Pre-Event Planet Dove



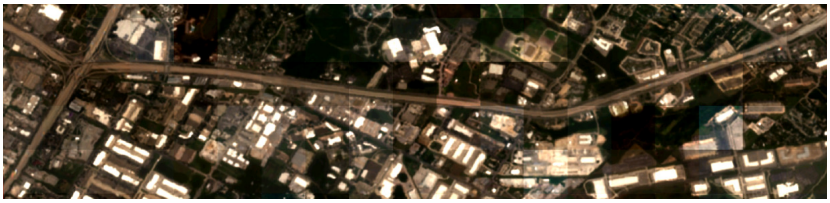
(e) Post-Event Planet Dove

(f) Change Detection Ground Truth

Fig. 6: Figures of the region of interest in the port of Beirut.

the quality of the figures, Figure 5a clearly shows great contrast between features due to it coming from a higher spatial resolution satellite when compared to Figure 5b. Figure 5c is the synthetic high-resolution image generated with the proposed model using Figure 5b as input. It exhibits a higher contrast between features, compared to the Sentinel-II version, while maintaining feature consistency between patches to the point that it becomes almost impossible to distinguish individual patches in the full images. The same can be said about Figure 5f, which shows a close-up view of a region located around the center of Figure 5c, where homogeneity is still observed, and the super-resolution aspect of the method starts to be noticeable. That is, high-frequency information not clearly observable in Figure 5e gets enhanced by the algorithm, resulting in Figure 5f. Finally, the last row depicts a highly zoomed-up area of the city, where individual residential buildings can be visualized. It is clearly noticeable in Figure 5g how the higher spatial resolution information provided by the Planet Dove sensor enables the visualization of details otherwise imperceptible in the Sentinel-II image (Figure 5h), as well as properly separates features coming from different elements in the image, e.g., roads are easily distinguishable from buildings, whereas in 5h the pixels from narrow roads are interleaved with pixels from buildings. This drawback of Sentinel-II images is alleviated by the proposed method, as can be observed in Figure 5i. There, pixel-level details are almost

(a) Pre-Event Sentinel-II



(b) Pre-Event Synthetic Planet Dove (Regression + $\mathbf{W}$)

(c) Pre-Event Synthetic Planet Dove (Proposed)

(d) Pre-Event Planet Dove

(e) Post-Event Planet Dove

(f) Change Detection Ground Truth

Fig. 7: Figures of the region of interest in Austin.

as clear and as distinguishable as in Figure 5g. In addition to this, the synthetic image does not show visible undesirable hallucinations, which enables change detection tasks.

Figures 6 and 7 respectively show images from the selected regions of Beirut and Austin. Their first, second, third and fourth rows individually show pre-event Sentinel-II, pre-event Synthetic Planet Dove (Regression + $\mathbf{W}$), pre-event Synthetic Planet Dove (Proposed), pre-event Planet Dove, post-event Planet Dove, and the change detection ground truth obtained by applying Algorithm 5, which is discussed in Section V-B, using Planet Dove pre and post event images. Instead of manual annotations, we use this ground truth since our I2I translation method aims to closely match low-resolution input to the style and resolution of high-resolution images, specifically Planet Dove in our experiments. In both regions, the tonality of the synthetic images generated using the proposed method (Figs. 6c and 7c) closely resemble those of the Planet pre-event images (Figs. 6d and 7d). Also, when they are compared with the pre-event Sentinel-II images used as inputs to the DDM-based method (Figs. 6a and 7a), it is easily observable how they have an improved apparent resolution and contrast. It is also noticeable how no perceptible undesired hallucinated features are present, favoring their use for change detection. Compared to the Regression + $\mathbf{W}$ synthetic images (Figs. 6b and 7b), our method delivered images with higher contrast and visual resolution, which translated into images that better match the characteristics of Planet Dove imagery.

Algorithm 5 Targetless Change Detection

Input: $\boldsymbol{I}_{\text{pre}}$, $\boldsymbol{I}_{\text{post}}$, l, n_{ch}, h, w, ω, w_{gauss}, w_{otsu}, e_{max}, n_{min}

Output: $\boldsymbol{I}_{\text{diff}}$

1: $\boldsymbol{I}_{\text{pre}} \leftarrow \left[\min\left(I_{\text{pre}}^{i,j,k}, \mu\left(\boldsymbol{I}_{\text{pre}}\right) + 6\sigma_{\boldsymbol{I}_{\text{pre}}}\right)\right]_{1\leq i\leq n_{\text{ch}},\ ,1\leq j\leq h,\ 1\leq k\leq w}$

2: $\boldsymbol{I}_{\text{post}} \leftarrow \left[\min\left(I_{\text{post}}^{i,j,k}, \mu\left(\boldsymbol{I}_{\text{post}}\right) + 6\sigma_{\boldsymbol{I}_{\text{post}}}\right)\right]_{1\leq i\leq n_{\text{ch}},\ ,1\leq j\leq h,\ 1\leq k\leq w}$

3: $\boldsymbol{I}_{\text{pre}} \leftarrow \text{GaussianBlur}\left(\frac{\boldsymbol{I}_{\text{pre}} - \min \boldsymbol{I}_{\text{pre}} \cdot \mathbf{J}_{n_{\text{ch}}\times h\times w}}{\max \boldsymbol{I}_{\text{pre}} - \min \boldsymbol{I}_{\text{pre}}}, w_{\text{gauss}}\right)$

4: $\boldsymbol{I}_{\text{post}} \leftarrow \text{GaussianBlur}\left(\frac{\boldsymbol{I}_{\text{post}} - \min \boldsymbol{I}_{\text{post}} \cdot \mathbf{J}_{n_{\text{ch}}\times h\times w}}{\max \boldsymbol{I}_{\text{post}} - \min \boldsymbol{I}_{\text{post}}}, w_{\text{gauss}}\right)$

5: $\boldsymbol{I}_{\text{pre}} \leftarrow \frac{\boldsymbol{I}_{\text{pre}} - \mu(\boldsymbol{I}_{\text{pre}})\mathbf{J}_{n_{\text{ch}}\times h\times w}}{\sigma(\boldsymbol{I}_{\text{pre}})}$

6: $\boldsymbol{I}_{\text{post}} \leftarrow \frac{\boldsymbol{I}_{\text{post}} - \mu(\boldsymbol{I}_{\text{post}})\mathbf{J}_{n_{\text{ch}}\times h\times w}}{\sigma(\boldsymbol{I}_{\text{post}})}$

7: $\boldsymbol{I}_{\text{diff}} \leftarrow \left(\boldsymbol{I}_{\text{post}} - \boldsymbol{I}_{\text{pre}}\right)^{\odot 2}$

8: $\boldsymbol{I}_{\text{diff}} \leftarrow \frac{1}{n_{\text{ch}}} \sum_{i=1}^{n_{\text{ch}}} \boldsymbol{I}_{\text{diff}}^{i}$

9: $\boldsymbol{I}_{\text{diff}} \leftarrow \frac{\boldsymbol{I}_{\text{diff}} - \min \boldsymbol{I}_{\text{diff}} \cdot \mathbf{J}_{n_{\text{ch}}\times h\times w}}{\max \boldsymbol{I}_{\text{diff}} - \min \boldsymbol{I}_{\text{diff}}}$

10: $\boldsymbol{I}_{\text{diff}} \leftarrow \mathbf{H}\left(\boldsymbol{I}_{\text{diff}} - \omega \mathbf{J}_{n_{\text{ch}}\times h\times w}\right) \odot \boldsymbol{I}_{\text{diff}}$

11: $\boldsymbol{I}_{\text{diff}} \leftarrow \text{Otsu}\left(\boldsymbol{I}_{\text{diff}}, w_{\text{otsu}}\right)$

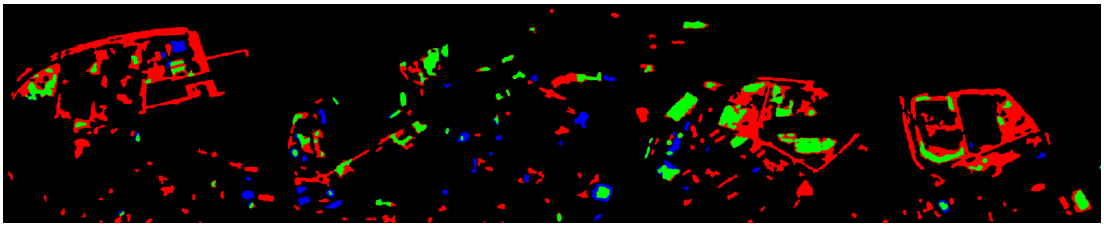
12: $\boldsymbol{L}_{\text{diff}} \leftarrow \text{DBSCAN}\left(\boldsymbol{I}_{\text{diff}}, e_{\text{max}}, n_{\text{min}}\right)$

13: $\boldsymbol{I}_{\text{diff}} \leftarrow \begin{cases} I_{\text{diff}}^{i,j,k} & \text{if } L_{\text{diff}}^{i,j,k} \neq -1 \\ 0 & \text{otherwise} \end{cases}$
$1 \leq i \leq n_{\text{ch}},\ ,1 \leq j \leq h,\ 1 \leq k \leq w$

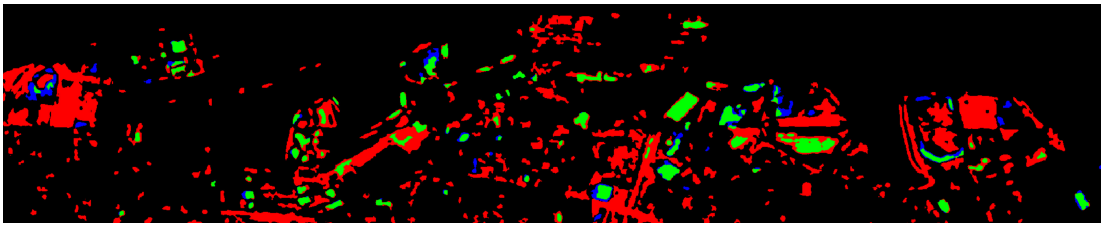
B. Change Detection

1) Procedure: Algorithm 5 defines a simple change detection procedure for a pair of images, $\mathbf{I}_{\text{pre}}$ and $\mathbf{I}_{\text{post}}$, taken before and after a change event, respectively. It is targetless, meaning that no specific class of detections is to be highlighted, but instead, all changes are treated equally. The algorithm, which is used in the change detection experiments presented henceforth, is a structured approach to targetless change detection, leveraging the robustness of Otsu's method [36], a well-known adaptive thresholding technique. It works by automatically selecting the optimal threshold value to separate pixels into two classes, foreground, and background, based on their grayscale levels, in a sliding-window fashion. Intuitively, it calculates the threshold that minimizes intra-class variance, effectively distinguishing between changed and unchanged areas in an image.

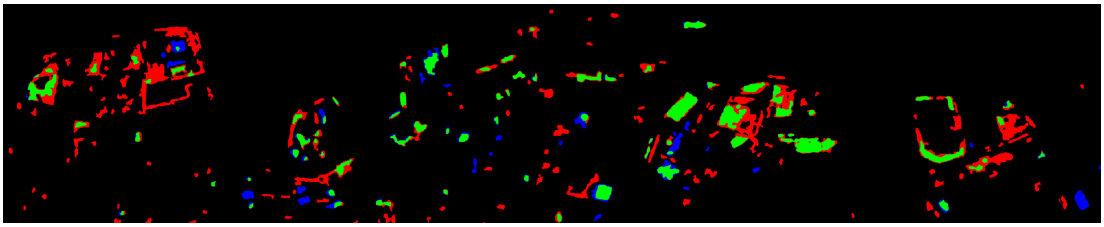
Lines 1 and 2 of Algorithm 5 individually clip $\mathbf{I}_{\text{pre}}$ and $\mathbf{I}_{\text{post}}$ to their mean plus six times their standard deviations, to remove extreme outliers that could affect the difference image calculation. Lines 3 and 4 apply Gaussian filtering to $[0, 1]$-normalized versions of the images, to smooth the images, which alleviates the value transitions, suppressing potential false alarms. Following, in Lines 5 and 6, the images are standardized so that the pixel values of both images become comparable. Line 7 computes the squared difference between the two images in order to highlight the changes, followed by a channel-wise mean in Line 8, to compact the change map into a 2D matrix. Line 10 zeroes out elements of the difference image that are smaller than ω. $\mathbf{H}$ is the element-wise Heaviside step function. By setting such a global threshold, we effectively filter out minor variations in pixel values that are likely due to noise. This pre-processing step simplifies the image data, ensuring that when Otsu's method is applied, it can more accurately focus on the substantial changes—those that are of real interest in the context of the event being analyzed. Then, in Line 11, the Otsu's adaptive thresholding [36] is applied with window size of w_{otsu}, to generate a binary map whose nonzero elements correspond to potential changes. In Lines 12 and 13, to alleviate the presence of detection noise, i.e., to remove positive pixels that are completely isolated and scattered through the binary map, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [37] algorithm is applied. It locates clusters of positive pixels of significant size and discards pixels that do not belong to any detected cluster, classifying them hereby as noisy pixels. Its parameters are the maximum Euclidean distance between pixels of a cluster, e_{max}, and the minimum number of pixels that a cluster should contain, n_{min}. The output of Line 12, $\boldsymbol{L}_{\text{diff}}$, is a matrix with the dimensions of I_{diff}, whose pixels are integers from -1 to the number of identified clusters. All pixels equal to -1 are classified as noise and are finally discarded in the last line of Algorithm 5, resulting in a clean binary map, where nonzero pixels are classified as changed pixels. For the executed experiments, the values of w_{gauss}, w_{otsu}, e_{max}, and n_{min} were set to 11, 1023, 5, and 48, respectively.



(a) Sentinel-II (Pre) + Planet (Post)



(b) Regression + $\mathbf{W}$ Synthetic Planet (Pre) + Planet (Post)



(c) Proposed Synthetic Planet (Pre) + Planet (Post)

(d) Ground Truth—Planet Pre and Post

Fig. 8: Change detection results in Beirut Port Region for a fixed detection rate of 0.75. Pixels in green, blue, and red are true positives, false negatives, and false positives, respectively.

2) Experiments: Two pairs of Sentinel-II pre-event and Planet Dove post-event images from two different cities were selected for a change detection comparison experiment. We also use their respective pre-event Planet Dove images to produce a change detection ground truth. The Sentinel-II pre-event images are part of the test set mentioned in Section IV. The first pair of images is in Beirut, Lebanon, surrounding the port region. It has been extracted from the northern part of the images in Figure 1, where no clouds are present. This image has been considered because it showcases multiple different types of changes, making targetless change detection useful. Moreover, it is a good example to test the robustness of the proposed method. The second pair of images is in Austin, USA, where many construction and terrain-related changes are present. The metrics used henceforth are the detection rate (DR) (recall) and the false alarm rate (FAR), which is defined as the ratio between the number of pixels misclassified as changes and the total number of unchanged pixels. These are metrics that properly quantify the trade-off between over and under detection.

Using the images presented in Figures 6 and 7, quantitative change detection results were obtained and presented in Figures 8 and 10 for Beirut and Austin, respectively. Pixels in green correspond to true positives, in red to false positives, and in blue to false negatives. The first, second, and third rows contain, individually, change detection results generated using as pre-event images: a Sentinel-II image (Figs. 8a and 10a), a synthetic Planet Dove image using the Regression + $\mathbf{W}$ model (Figs. 8b and 10b), described in the previous section, and a synthetic Planet Dove image generated with the proposed DDM-based algorithm (Figs. 8c and 10c). The Coloring procedure executed in the generation of the synthetic images made use of color information extracted from the post-Planet Dove image patches. A threshold ω has been chosen for each generated change map such that $\mathrm{DR} = 0.75$, which is an operating point that highlights the effects of the proposed algorithm.

Comparing the results in Fig. 8, the proposed method has

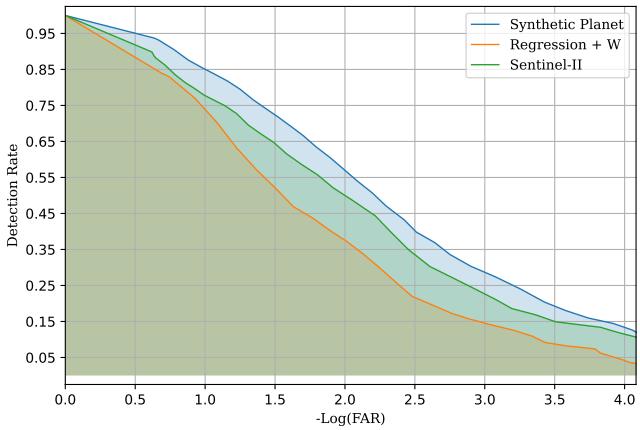
been able to heavily suppress false alarms. In fact, the change maps for Sentinel-II and Regression + **W** as pre-images show a staggering amount of false alarms, greatly surpassing the number of true positives. When the proposed method is applied, however, the majority of such false alarms are completely avoided. This visual superiority is further confirmed by the calculated metrics. For the Beirut region in Fig. 8, with DR $= 0.75$, the proposed method reached a FAR of 0.03933 ($-log_{10}(\text{FAR}) = -1.405$), whereas when using the Sentinel-II pre-event image, and the Regression + **W** synthetic pre-event image, the FAR was of 0.0731 ($-log_{10}(\text{FAR}) = -1.136$) and 0.1055 ($-log_{10}(\text{FAR}) = -0.9769$), respectively.

For the Austin region in Fig. 10, the improvements are less visually evident but still numerically expressive. In Fig. 10a, when the Sentinel-II image is used as pre-event image, multiple false alarms related to unchanged buildings can be spotted, specially in the bottom of the image, which are not observed in the change map for the proposed method's change map (Fig. 10c). Moreover, for the regression-based change detection in Fig. 10b, a big false alarm is spotted in the eastern part of the image, which is not present in the change map of the proposed method. The metrics again prove the superiority of our technique: for DR $= 0.75$, using the synthetic image generated by the proposed algorithm as the pre-event image leads to a FAR of 0.001678 ($-log_{10}(\text{FAR}) = -2.775$), whereas when using Sentinel-II leads to a FAR of 0.007194 ($-log_{10}(\text{FAR}) = -2.143$), and to a FAR of 0.003648 ($-log_{10}(\text{FAR}) = -2.438$) when Regression + **W** is used.

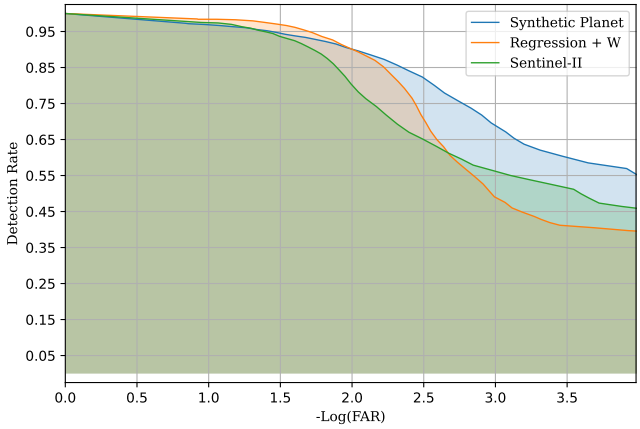
To better evaluate the performance of the proposed algorithm for all possible choices of clipping threshold ω, employed in Line 10 of Algorithm 5, Figures 9a and 9b present receiving operating characteristic (ROC) comparisons between the baseline change detection and change detection using the proposed method, for Beirut port and Austin regions, respectively. These curves were obtained by varying ω between 0 and 1.0, with steps of 0.01. As can be visualized in the curves, ω adjusts how permissive or restrictive should the algorithm be towards potential changes: a trade-off between higher detection rates and lower false alarm rates.

In Fig. 9a, the proposed method beats both compared methods for all choices of ω, which displays how it has been able to significantly reduce false alarms for a desired detection rate. For fixed detection rates, the difference between the false alarm rate of the proposed method and the other two cases is up to 0.7 in the logarithmic scale or 5 times lower in the linear scale (DR=0.45).

Meanwhile, in Fig. 9b, the performance of the proposed method is on par with the other two methods up to $-log_{10}(\text{FAR}) = 2$, displaying a slightly lower DR compared to the regression-based model: a maximum difference of 0.01 in the DR. After this point, however, the proposed method beats by a large margin the other two approaches, reaching a FAR difference of up to 1.25 in the logarithmic scale, or 17.8 times lower in the linear scale.



(a) Beirut Port Region

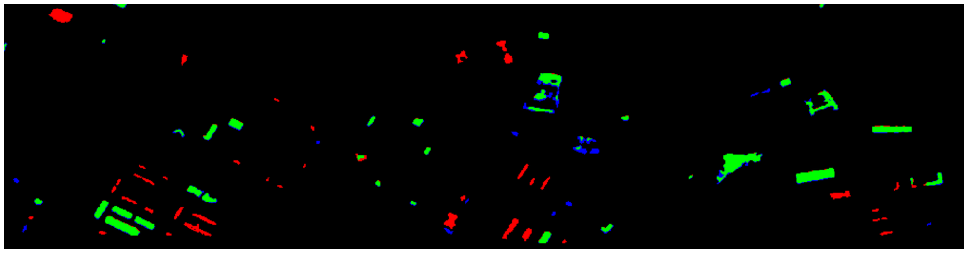


(b) Austin Region

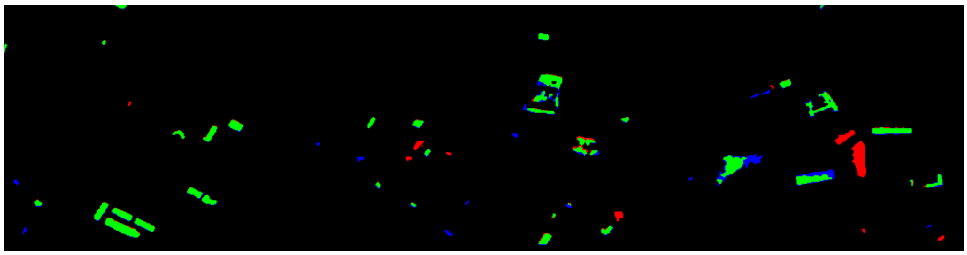
Fig. 9: Performance comparison of heterogeneous change detection region using a Planet Dove frame as a post-event image. *Sentinel-II* refers to when a Sentinel-II image is directly used as a pre-event image. *Regression + W* refers to when using the regression-based model with the whitening operation to generate a synthetic pre-event image. *Synthetic Planet* refers to when a synthetic Planet Dove image generated by the proposed I2I method is used as a pre-event image.

VI. CONCLUSION

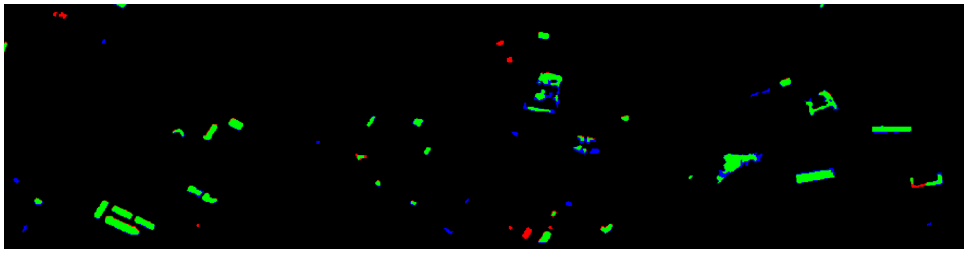
In this article, we introduced an innovative deep learning-based method that uses denoising diffusion models to generate synthetic high spatial resolution satellite images following the characteristics of a high-resolution optical sensor, using low spatial resolution images from another sensor. We trained and tested our method on a large and diverse data set of paired images from these sensors and compared it with other solutions, including the popular classifier-free guided DDIM framework. The success of our method can be attributed to a novel training regimen that deviates from traditional Denoising Diffusion Probabilistic Models. By incorporating color standardization and initial noise selection techniques,



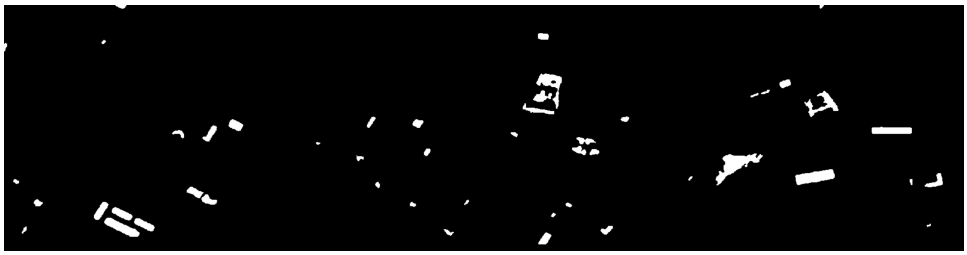
(a) Sentinel-II (Pre) + Planet (Post)



(b) Regression + $\mathbf{W}$ Synthetic Planet (Pre) + Planet (Post)



(c) Proposed Synthetic Planet (Pre) + Planet (Post)



(d) Ground Truth—Planet Pre and Post

Fig. 10: Change detection results in Austin Region for a fixed detection rate of 0.75. Pixels in green, blue, and red are true positives, false negatives, and false positives, respectively.

we have achieved a significant alignment between input and output patches, thereby reducing the likelihood of model hallucinations. These strategic modifications are important to ensure that the diffusion network is not misled by the high variability in patch coloration found within large training datasets. We also applied our method to the task of heterogeneous change detection (HCD) between images from different sensors and demonstrated that our method can improve the performance of a targetless change detection algorithm in two urban areas: Beirut, Lebanon, and Austin, USA. Our method can be useful for HCD applications, especially in cases where the availability of high-resolution images from the same sensor is limited or costly. In future work, we plan to extend our method to other combinations of sensors and investigate the employment of synthetic images in target change detection.

References

[1] C. Saharia, W. Chan, H. Chang, *et al.*, "Palette: Image-to-Image Diffusion Models," *arXiv*, 2022.

[2] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *arXiv*, 2021.

[3] M. Seo, Y. Oh, D. Kim, D. Kang, and Y. Choi, "Improved Flood Insights: Diffusion-Based SAR to EO Image Translation," *arXiv*, 2023.

[4] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic Image Synthesis with Spatially-Adaptive Normalization," *arXiv*, 2019.

[5] S. F. Ismael, K. Kayabol, and E. Aptoula, "Unsupervised Domain Adaptation for the Semantic Segmentation of Remote Sensing Images via One-Shot Image-to-Image Translation," *IEEE Geoscience and Remote*

Sensing Letters, vol. 20, 2023. DOI: 10.1109/LGRS.2023.3281458.

[6] O. Tasar, S. L. Happy, Y. Tarabalka, and P. Alliez, "SEMI2I: Semantically Consistent Image-to-Image Translation for Domain Adaptation of Remote Sensing Data," Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 1837–1840, ISBN: 9781728163741. DOI: 10.1109/IGARSS39084.2020.9323711.

[7] M. Zhang, J. Xu, C. He, W. Shang, Y. Li, and X. Gao, "Sar-to-optical image translation via thermodynamics-inspired network," 2023.

[8] M. Sokolov, C. Henry, J. Storie, C. Storie, V. Alhassan, and M. Turgeon-Pelchat, "High-Resolution Semantically Consistent Image-to-Image Translation," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 16, pp. 482–492, 2023. DOI: 10.1109/JSTARS.2022.3226705.

[9] B. Fang, G. Chen, R. Kou, M. E. Paoletti, J. M. Haut, and A. Plaza, "CIT: Content-invariant translation with hybrid attention mechanism for unsupervised change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 321–339, Oct. 2023. DOI: 10.1016/j.isprsjprs.2023.09.012.

[10] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (gan) based change detection network for optical and sar remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 179, pp. 14–34, 2021. DOI: 10.1016/j.isprsjprs.2021.07.007.

[11] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023. DOI: 10.1109/TGRS.2023.3341437.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *arXiv*, 2020.

[13] H. Mansourifar, A. Moskovitz, B. Klingensmith, D. Mintas, and S. J. Simske, "GAN-Based Satellite Imaging: A Survey on Techniques and Applications," *IEEE Access*, vol. 10, pp. 118 123–118 140, 2022. DOI: 10.1109/ACCESS.2022.3221123.

[14] J. Ho and T. Salimans, "Classifier-Free Diffusion Guidance," *arXiv*, 2022.

[15] Jiaming Song and Chenlin Meng and Stefano Ermon, "Denoising Diffusion Implicit Models," *arXiv*, 2022.

[16] A. Nichol and P. Dhariwal, "Improved Denoising Diffusion Probabilistic Models," *arXiv*, 2021.

[17] W. Wang, J. Bao, W. Zhou, *et al.*, "Semantic Image Synthesis via Diffusion Models," *arXiv*, 2022.

[18] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *arXiv*, 2018.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017.

[20] A. Horé and D. Ziou, "Image Quality Metrics: PSNR vs. SSIM," in *20th International Conference on Pattern Recognition*, 2010, pp. 2366–2369. DOI: 10.1109/ICPR.2010.579.

[21] H. Chen, N. Yokoya, and M. Chini, "Fourier domain structural relationship analysis for unsupervised multimodal change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 198, pp. 99–114, 2023, ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2023.03.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092427162300062X.

[22] R. Touati, M. Mignotte, and M. Dahmane, "Anomaly Feature Learning for Unsupervised Change Detection in Heterogeneous Images: A Deep Sparse Residual Model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 588–600, 2020. DOI: 10.1109/JSTARS.2020.2964409.

[23] C. Zhang, Y. Feng, L. Hu, *et al.*, "A domain adaptation neural network for change detection with heterogeneous optical and SAR remote sensing images," *International Journal of Applied Earth Observation and Geoinformation*, vol. 109, May 2022. DOI: 10.1016/j.jag.2022.102769.

[24] J.-J. Wang, N. Dobigeon, M. Chabert, D.-C. Wang, T.-Z. Huang, and J. Huang, "CD-GAN: a robust fusion-based generative adversarial network for unsupervised remote sensing change detection with heterogeneous sensors," *arXiv*, 2023.

[25] Z. Y. Lv, H. T. Huang, X. Li, *et al.*, "Land Cover Change Detection with Heterogeneous Remote Sensing Images: Review, Progress, and Perspective," *Proceedings of the IEEE*, vol. 110, pp. 1976–1991, 12 Dec. 2022. DOI: 10.1109/JPROC.2022.3219376.

[26] Y. Wu, J. Li, Y. Yuan, A. K. Qin, Q. G. Miao, and M. G. Gong, "Commonality Autoencoder: Learning Common Features for Change Detection From Heterogeneous Images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, pp. 4257–4270, 9 Sep. 2022. DOI: 10.1109/TNNLS.2021.3056238.

[27] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image Super-Resolution via Iterative Refinement," *arXiv*, 2021.

[28] G. Daras, Y. Dagan, A. G. Dimakis, and C. Daskalakis, "Consistent Diffusion Models: Mitigating Sampling Drift by Learning to be Consistent," *arXiv*, 2023.

[29] G. P. Meyer, "An Alternative Probabilistic Interpretation of the Huber Loss," *arXiv*, 2020.

[30] L. Liu, H. Jiang, P. He, *et al.*, "On the Variance of the Adaptive Learning Rate and Beyond," *arXiv*, 2021.

[31] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging Weights Leads to Wider Optima and Better Generalization," *arXiv*, 2019.

[32] T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson, "Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs," *arXiv*, 2018.

[33] T. Hang, S. Gu, C. Li, *et al.*, "Efficient Diffusion Training via Min-SNR Weighting Strategy," *arXiv*, 2023.

[34] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric," *arXiv*, 2018.

[35] S. Lathuiliere, P. Mesejo, X. Alameda-Pineda, and R. Horaud, "A comprehensive analysis of deep regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 9, pp. 2065–2081, Sep. 2020. DOI: 10.1109/tpami.2019.2910523.

[36] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979. DOI: 10.1109/TSMC.1979.4310076.

[37] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, pp. 226–231.