# StegoGAN: Leveraging Steganography for Non-Bijective Image-to-Image Translation

Sidi Wu [1*] Yizi Chen [1*] Samuel Mermet [2] Lorenz Hurni [1]
Konrad Schindler [1] Nicolas Gonthier [3,2] Loic Landrieu [4,2]

[1] ETH Zurich [2] Univ Gustave Eiffel, IGN, ENSG, LASTIG [3] IGN [4] Univ Gustave Eiffel, CNRS, Ecole des Ponts, LIGM

## Abstract

*Most image-to-image translation models postulate that a unique correspondence exists between the semantic classes of the source and target domains. However, this assumption does not always hold in real-world scenarios due to divergent distributions, different class sets, and asymmetrical information representation. As conventional GANs attempt to generate images that match the distribution of the target domain, they may hallucinate spurious instances of classes absent from the source domain, thereby diminishing the usefulness and reliability of translated images. CycleGAN-based methods are also known to hide the mismatched information in the generated images to bypass cycle consistency objectives, a process known as steganography. In response to the challenge of non-bijective image translation, we introduce StegoGAN, a novel model that leverages steganography to prevent spurious features in generated images. Our approach enhances the semantic consistency of the translated images without requiring additional postprocessing or supervision. Our experimental evaluations demonstrate that StegoGAN outperforms existing GAN-based models across various non-bijective image-to-image translation tasks, both qualitatively and quantitatively. Our code and pretrained models are accessible at* https://github.com/sian-wusidi/StegoGAN.

## 1. Introduction

Image-to-image translation is an active research subject with impactful applications ranging from changing the style of images [12, 36] to automatically creating maps from satellite images [36] or changing the modality of medical images [9]. When the source and target domains exhibit substantial differences, ensuring the semantic consistency between input images and their translation becomes particularly challenging [8]. Our work explores the surprisingly uncharted field of adversarial, non-bijective image-to-image translation.

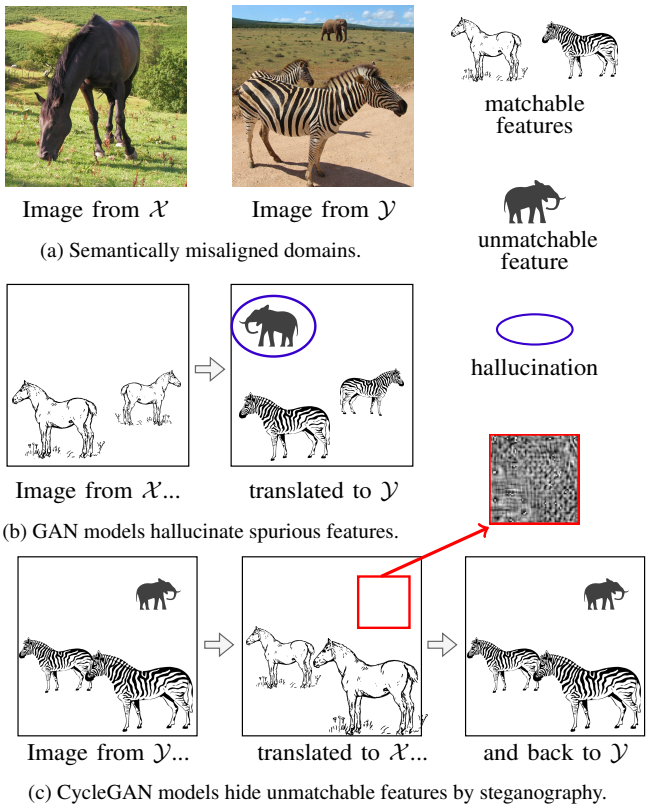


Figure 1. **Non-Bijective Translation.** When image domains present classes without equivalence (a), GAN models tend to hallucinate spurious features when translating images (b). A related phenomenon is steganography, where CycleGAN-based models covertly encode features in low-amplitude patterns to bypass cycle consistency (c). Instead of disabling this phenomenon, we harness steganography to prevent the hallucination of spurious features.

**Non-Bijective Image Translation.** Existing translation methods assume a one-to-one correspondence between classes of the source and target domains: horses to zebras [36], satellite image features to their cartographic repre-

*Equal contribution.

sentation [6, 36], or distinct cell types viewed under varying medical imaging modalities [9]. However, as illustrated in Figure 1a, this assumption does not always hold. For instance, when considering a dataset with horses and one with zebras in their habitat, zebra images may include background elements with no equivalent in the source domain—like elephants. Similarly, in map translation toponyms (*i.e.*, place names printed onto the map) do not have counterparts in satellite images [6] . We qualify classes of the target domain without equivalent in the source domain as *unmatchable*.

As illustrated in Figure 1b, by trying to reproduce the distribution of the target domain, GANs may hallucinate *spurious* features or textures in the generated images, *i.e.*, objects without equivalent in the source image. This is particularly frequent for unmatchable classes [8]. While this can be perfectly acceptable for some applications [23], adding nonexistent tumors in MRI scans or incorrect toponyms in maps can severely degrade the usefulness of the translation result. Instead of detecting and removing these artifacts in post-processing, we propose an approach that directly prevents their generation using steganography.

**GAN Steganography.** To ensure its semantic consistency, CycleGAN [36] back-translates the generated image to the source image. However, unmatchable classes of the source domain cannot be encoded into meaningful features in the images generated in the target domain. As shown in Figure 1c, these models can instead *cheat* by encoding the necessary information into quasi-invisible patterns in the generated images [7]. This process, known as steganography, allows GANs to perform seemingly impossible back-translation. For instance, in a map-to-satellite task, the model can restore the correct names of towns from satellite images that appear visually correct. This phenomenon is often viewed as a quirky optimization flaw, easily fixable by adding noise or blur [11, 20, 26].

**StegoGAN.** We propose StegoGAN, a model that leverages steganography to detect and mitigate semantic misalignment between domains. In settings where the domain mapping is non-bijective, StegoGAN experimentally demonstrates superior semantic consistency over other GAN-based models both visually and quantitatively, without requiring detection or inpainting steps. In addition, we publish three datasets from open-access sources as a benchmark for evaluating non-bijective image translation models.

## 2. Related Work

**GAN-Based Image Translation.** GAN-based image translation models transfer the style of images between domains with an adversarial perceptual loss [13]. When pairs of aligned images from both domains are available, the translated images can also be supervised by their fidelity with target images [18]. In practice, such pairs are not always available or even possible to obtain. In the absence of explicit equivalence between images, preserving the semantics of the input in the generated image is a priority. Multiple approaches have been proposed to address this challenge, such as density-based regularization [33], spatial mutual information [20, 26, 30], or cycle consistency losses [17, 21, 36].

**Asymmetric Image Translation.** Translating between domains with different semantic distributions is challenging. Existing approaches include focusing the network's attention on the most discriminative part of the input image [29], augmenting the consistency loss with geometric transformations [11], replacing the consistency reconstruction term with a contrastive loss [20, 26], or ensuring that the translation is robust to small perturbations of the input [19]. However, these methods assume a bijective relationship between the classes of the source and target domains.

Closest to our work is the model of Li *et al*. [23], which uses an auxiliary variable to model the information loss from information-rich domains (such as natural images) to information-poor domains (such as label maps). In turn, they use this variable to create realistic poor-to-rich domain translations. Our work differs as we precisely want to avoid the creation of spurious—albeit realistic—details when translating to a domain with unmatchable classes.

**CycleGAN Steganography.** Chu *et al*. [7] discovered that, when faced with unmatchable classes, CycleGAN [36] hides information in low-amplitude and high-frequency signals. The model uses these visually imperceptible patterns to recreate the source image and bypass the cycle loss. This contradicts the intention of the cycle consistency loss and makes the model more vulnerable to adversarial attacks [7]. Luckily, multiple approaches can prevent steganography, such as blurring [11], compressing [10], or adding noise [20] to the generated source images in the back-translation. Alternatively, the back-translation from poor to rich domains can be omitted in the cycle consistency loss [27, 35].

While steganography in CycleGAN can be problematic, it also offers an opportunity to analyse distribution differences. In StegAnomaly [3], a model is trained to translate healthy brain scans into a low-entropy domain with cycle consistency. When removing high-frequency components, the model error reveals anomalous structures. This approach, like ours, harnesses steganography for insightful domain analysis, albeit with a different goal.

## 3. Methods

We consider two image domains $\mathcal{X}$ and $\mathcal{Y}$ with respective semantic class sets $\mathcal{K}_\mathcal{X}$ and $\mathcal{K}_\mathcal{Y}$. The domains $\mathcal{X}$ and $\mathcal{Y}$ are considered bijective if there exists a function $\phi$ from $\mathcal{K}_\mathcal{X}$ to $\mathcal{K}_\mathcal{Y}$ such that each class $k_\mathcal{X}$ has a unique and natural semantically equivalent class $\phi(k_\mathcal{X})$ in $\mathcal{Y}$, and vice-versa. A class of $\mathcal{K}_\mathcal{Y}$ is said to be *unmatchable* if it doesn't have an

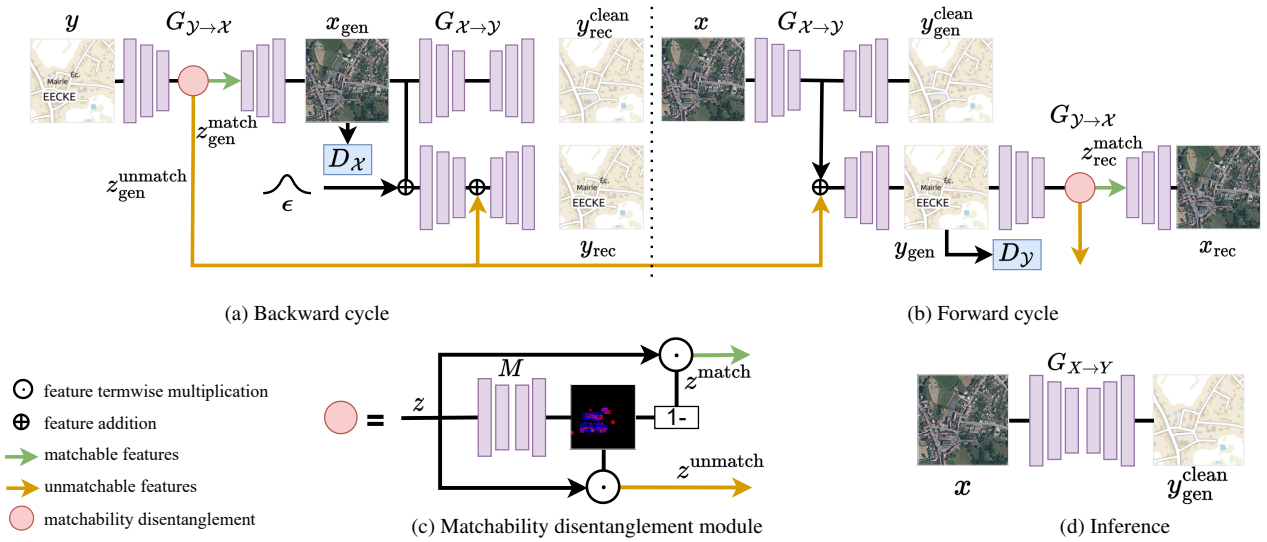


Figure 2. **Architecture.** To avoid spurious generation of unmatchable classes in non-bijective image translation, we propose to make the steganographic process explicit and in feature-space. Our model runs the backward cycle first (a), then the forward translation cycle (b). Thanks to our matchability disentanglement module (c), we can separate the matchable and unmatchable information while translating images from domain $\mathcal{Y}$ to $\mathcal{X}$. We can then produce generated and reconstructed images with and without unmatchable features. At inference time (d), our model operates like a normal image translation model.

equivalent in $\mathcal{K}_{\mathcal{X}}$. While this notion is somewhat subjective, many applications have obvious examples: toponyms are unmatchable in satellite images, and tumors in scans of healthy patients.

**Objective.** Our goal is to learn a mapping $G : \mathcal{X} \mapsto \mathcal{Y}$ such that the translation $G(x)$ of any image $x \in \mathcal{X}$ aligns stylistically with images from $\mathcal{Y}$, while preserving the semantic content of $x$. If $\mathcal{K}_{\mathcal{Y}}$ contains an unmatchable class $k_{\mathcal{Y}}^{\text{unmatch}}$, images translated from $\mathcal{X}$ to $\mathcal{Y}$ should not contain any instances of $k_{\mathcal{Y}}^{\text{unmatch}}$. However, in an attempt to match the distribution of $\mathcal{Y}$, GAN models often create spurious instances of unmatchable classes in their translated images. In this paper, we propose a method that employs steganography to prevent the generation of such spurious information.

### 3.1. CycleGAN

CycleGAN [36] learns unpaired image translation by enforcing the consistency between the input image and its back-translation from the generated image. It uses two generators $G_{\mathcal{X}\mapsto\mathcal{Y}} : \mathcal{X} \mapsto \mathcal{Y}$ and $G_{\mathcal{Y}\mapsto\mathcal{X}} : \mathcal{Y} \mapsto \mathcal{X}$, and two domain discriminators $D_{\mathcal{X}} : \mathcal{X} \mapsto \{0, 1\}$, $D_{\mathcal{Y}} : \mathcal{Y} \mapsto \{0, 1\}$ which predict whether a sample is generated (0) or real (1).

In the following, when considering images $x \in \mathcal{X}$ or $y \in \mathcal{Y}$, we define the following short-hands: $x_{\text{gen}} := G_{\mathcal{Y}\mapsto\mathcal{X}}(y)$ and $y_{\text{gen}} := G_{\mathcal{X}\mapsto\mathcal{Y}}(x)$ for the generated images, and $x_{\text{rec}} := G_{\mathcal{Y}\mapsto\mathcal{X}}(y_{\text{gen}})$ and $y_{\text{rec}} := G_{\mathcal{X}\mapsto\mathcal{Y}}(x_{\text{gen}})$ for the reconstructed images. We now detail the losses of CycleGAN.

**Adversarial loss.** The adversarial loss [13] encourages the discriminators $D_{\mathcal{X}}$ and $D_{\mathcal{Y}}$ to distinguish between authentic and generated images, while pushing the generators $G_{\mathcal{X}\mapsto\mathcal{Y}}$ and $G_{\mathcal{Y}\mapsto\mathcal{X}}$ to create credible images:

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = &\mathbb{E}_{y\sim\mathcal{Y}} \log(D_{\mathcal{Y}}(y)) + \mathbb{E}_{x\sim\mathcal{X}} \log(1 - D_{\mathcal{Y}}(y_{\text{gen}})) \\ &+ \mathbb{E}_{x\sim\mathcal{X}} \log(D_{\mathcal{X}}(x)) + \mathbb{E}_{y\sim\mathcal{Y}} \log(1 - D_{\mathcal{X}}(x_{\text{gen}})). \quad (1)\end{aligned}$$

**Cycle consistency.** The cycle consistency loss ensures that the back-translation of $y_{\text{gen}}$ to domain $\mathcal{X}$ is close to the original image $x$, and likewise for $x_{\text{gen}}$ and $y$:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x\sim\mathcal{X}} \|x_{\text{rec}} - x\| + \mathbb{E}_{y\sim\mathcal{Y}} \|y_{\text{rec}} - y\| , \quad (2)$$

with $\|\cdot\|$ the pixel-wise $L_1$ norm.

**Identity loss.** The identity loss regularizes the generators to be close to identity, generally improving color composition:

$$\mathcal{L}_{\text{id}} = \mathbb{E}_{x\sim\mathcal{X}} \|G_{\mathcal{Y}\mapsto\mathcal{X}}(x) - x\| + \mathbb{E}_{y\sim\mathcal{Y}} \|G_{\mathcal{X}\mapsto\mathcal{Y}}(y) - y\|. \quad (3)$$

**Final Loss.** The final objectives are:

$$\mathcal{L}(G_{\mathcal{X}\mapsto\mathcal{Y}}, G_{\mathcal{Y}\mapsto\mathcal{X}}) = \mathcal{L}_{GAN} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{id}}\mathcal{L}_{\text{id}} , \quad (4)$$

$$\mathcal{L}(D_{\mathcal{X}}, D_{\mathcal{Y}}) = -\mathcal{L}_{\text{GAN}} , \quad (5)$$

with $\lambda_{\text{cyc}}$ and $\lambda_{\text{id}}$ non-negative hyperparameters.

### 3.2. StegoGAN

We introduce StegoGAN, a novel model building on the CycleGAN framework [36], designed specifically for scenarios where domains $\mathcal{X}$ and $\mathcal{Y}$ lack a bijective relationship. When generating images $y_{\text{gen}}$ in $\mathcal{Y}$, the generator $G_{\mathcal{X}\mapsto\mathcal{Y}}$ may add

spurious instances of an unmatchable class $k_{\mathcal{Y}}^{\text{unmatch}}$ in order to deceive the discriminator $D_{\mathcal{Y}}$. Our goal is to prevent the generation of such hallucinated instances. To achieve this, we leverage steganography to explicitly disentangle the matchable and unmatchable information in the backward cycle ($y \mapsto x_{\text{gen}} \mapsto y_{\text{rec}}$) and prevent the network from hallucinating in the forward translation ($x \mapsto y_{\text{gen}} \mapsto x_{\text{rec}}$). See Figure 2 for the overall design of our approach.

**Steganography.** In order to faithfully reconstruct $y$ with $y_{\text{rec}}$, the translated image $x_{\text{gen}}$ must somehow contain information about the instances of the unmatchable class $k_{\mathcal{Y}}^{\text{unmatch}}$. CycleGAN methods typically achieve this by hiding low-amplitude and high-frequency patterns in $x_{\text{gen}}$ that will be decoded by $G_{\mathcal{X}\mapsto\mathcal{Y}}$ and translated back to instances of $k_{\mathcal{Y}}^{\text{unmatch}}$ in $y_{\text{rec}}$. Adding low-intensity noise on each pixel is typically sufficient to destroy the hidden information and prevent steganography [7].

Steganography is often viewed as an optimization flaw that undermines the consistency loss. However, in the case of non-bijective translation, this is the only way to let $G_{\mathcal{X}\mapsto\mathcal{Y}}$ reconstruct instances of $k_{\mathcal{Y}}^{\text{unmatch}}$ in $y_{\text{rec}}$. Instead of disabling steganography, we propose to use it to our advantage to detect and prevent spurious generations. We adapt CycleGAN so that steganography takes place in feature-space instead of pixel-space, and in an explicit manner.

**Backward Cycle.** We decompose the generators $G_{\mathcal{X}\mapsto\mathcal{Y}}$ and $G_{\mathcal{Y}\mapsto\mathcal{X}}$ into two components: an encoder and a decoder, such that $G_{\mathcal{X}\mapsto\mathcal{Y}} = G^{\text{dec}}_{\mathcal{X}\mapsto\mathcal{Y}} \circ G^{\text{enc}}_{\mathcal{X}\mapsto\mathcal{Y}}$ and $G_{\mathcal{Y}\mapsto\mathcal{X}} = G^{\text{dec}}_{\mathcal{Y}\mapsto\mathcal{X}} \circ G^{\text{enc}}_{\mathcal{Y}\mapsto\mathcal{X}}$. The encoders map their inputs to feature maps of spatial dimension $H \times W$, where each pixel has $C$ channels. In the following, we denote the intermediary representation of $y$ in $G_{\mathcal{Y}\mapsto\mathcal{X}}$ by $z_{\text{gen}} := G^{\text{enc}}_{\mathcal{Y}\mapsto\mathcal{X}}(y)$. The feature map $z_{\text{gen}}$ encodes information about both matchable and unmatchable classes, which we want to disentangle.

We introduce a network $M\colon \mathbb{R}^{H\times W\times C} \mapsto [0,1]^{H\times W\times C}$ that assigns an *unmatchability* score between 0 and 1 to each pixel and channel. Here, a score of 1 indicates that the information does not have a counterpart in domain $\mathcal{X}$, while it does for 0. This process gives us the unmatchability mask $M(z_{\text{gen}})$, which we use to split $z_{\text{gen}}$ into its matchable and unmatchable parts:

$$z_{\text{gen}}^{\text{unmatch}} = M(z_{\text{gen}}) \odot z \tag{6}$$

$$z_{\text{gen}}^{\text{match}} = (1 - M(z_{\text{gen}})) \odot z \,, \tag{7}$$

with $\odot$ the pixel-wise and channel-wise Hadamard product. In our model, the generated image $x_{\text{gen}}$ is computed using only the matchable part of the representation:

$$x_{\text{gen}} = G^{\text{dec}}_{\mathcal{Y}\mapsto\mathcal{X}}(z_{\text{gen}}^{\text{match}}) \,. \tag{8}$$

We produce two reconstructions of $y$: $y_{\text{rec}}^{\text{clean}}$, which is a direct back-translation of $x_{\text{gen}}$ into the $\mathcal{Y}$ domain : $y_{\text{rec}}^{\text{clean}} = G_{\mathcal{X}\mapsto\mathcal{Y}}(x_{\text{gen}})$; and $y_{\text{rec}}$, which is generated by decoding a combination of the unmatchable part of $z_{\text{gen}}$ and the features extracted by $G^{\text{enc}}_{\mathcal{X}\mapsto\mathcal{Y}}$ from a noise-perturbed version of $x_{\text{gen}}$:

$$y_{\text{rec}} = G^{\text{dec}}_{\mathcal{X}\mapsto\mathcal{Y}}\left(G^{\text{enc}}_{\mathcal{X}\mapsto\mathcal{Y}}\left(x_{\text{gen}} + \epsilon\right) + z_{\text{gen}}^{\text{unmatch}}\right) \,, \tag{9}$$

with $\epsilon$ denoting random Gaussian noise of low amplitude applied to each pixel and channel of $x_{\text{gen}}$. This noise is added to destroy potential steganographic information in $x_{\text{gen}}$, therefore forcing $G_{\mathcal{X}\mapsto\mathcal{Y}}$ to rely only on $z_{\text{gen}}^{\text{unmatch}}$ to reconstruct unmatchable features in $y_{\text{rec}}$.

The key mechanisms to disentangle matchable and unmatchable information are twofold: (i) disturbing direct steganography with random noise, and (ii) explicitly providing unmatchable information to $G_{\mathcal{X}\mapsto\mathcal{Y}}$ *in feature-space*.

**Forward Cycle.** In the forward cycle $x \to y_{\text{gen}} \to x_{\text{rec}}$, the generator $G_{\mathcal{X}\mapsto\mathcal{Y}}$ may create spurious instances of unmatchable classes when translating $x$ to $\mathcal{Y}$ to fulfill the expectations of the discriminator $D_y$. To address this, we perform two distinct translations of $x$ in $\mathcal{Y}$: $y_{\text{gen}}$ has explicit access to the steganographic information $z_{\text{gen}}^{\text{unmatch}}$ extracted from the backward cycle, while $y_{\text{gen}}^{\text{clean}}$ does not:

$$y_{\text{gen}} = G^{\text{dec}}_{\mathcal{X}\mapsto\mathcal{Y}}\left(G^{\text{enc}}_{\mathcal{X}\mapsto\mathcal{Y}}(x) + z_{\text{gen}}^{\text{unmatch}}\right) \tag{10}$$

$$y_{\text{gen}}^{\text{clean}} = G_{\mathcal{X}\mapsto\mathcal{Y}}\left(x\right) \,. \tag{11}$$

The rationale is that $G^{\text{dec}}_{\mathcal{X}\mapsto\mathcal{Y}}$ has explicit access to information about the unmatchable classes of $y$, so it is not incentivized to invent them. For consistency with the backward step, where the decoder of $G_{\mathcal{Y}\mapsto\mathcal{X}}$ processes only matchable information as defined in (7), we use the same disentanglement approach for generating $x_{\text{rec}}$:

$$x_{\text{rec}} = G^{\text{dec}}_{\mathcal{Y}\mapsto\mathcal{X}}\left(\left(1 - M\left(z_{\text{rec}}\right)\right) \odot z_{\text{rec}}\right) \,, \tag{12}$$

where $z_{\text{rec}} = G^{\text{enc}}_{\mathcal{Y}\mapsto\mathcal{X}}(y_{\text{gen}})$ is the intermediary representation of $y_{\text{gen}}$ in the forward cycle.

**Mask Regularization.** To avoid degenerate behaviors of our explicit steganography mechanism, we enforce two priors on the unmatchability masks: (i) given that a well-posed translation problem predominantly involves matchable features, the masks should be sparse; (ii) to improve the model's interpretability, we favor mask values near 0 or 1, representing clear decisions about matchability. To enforce these priors, we regularize the masks with the non-convex $L_{0.5}$ norm [34]:

$$\mathcal{L}_{reg} = \mathbb{E}_{y\sim\mathcal{Y}}\|M(z_{\text{gen}})\|_{0.5} + \mathbb{E}_{x\sim\mathcal{X}}\|M(z_{\text{rec}})\|_{0.5} \,. \tag{13}$$

**Matchable Consistency.** The images $y_{\text{gen}}$ and $y_{\text{gen}}^{\text{clean}}$, as well as $y_{\text{rec}}$ and $y_{\text{rec}}^{\text{clean}}$, should be identical outside of unmatchable regions. To enforce this constraint, we design a function $I$ which takes an unmatchability mask $m$ as input, takes the channel-wise maximum values of all pixels, flips its

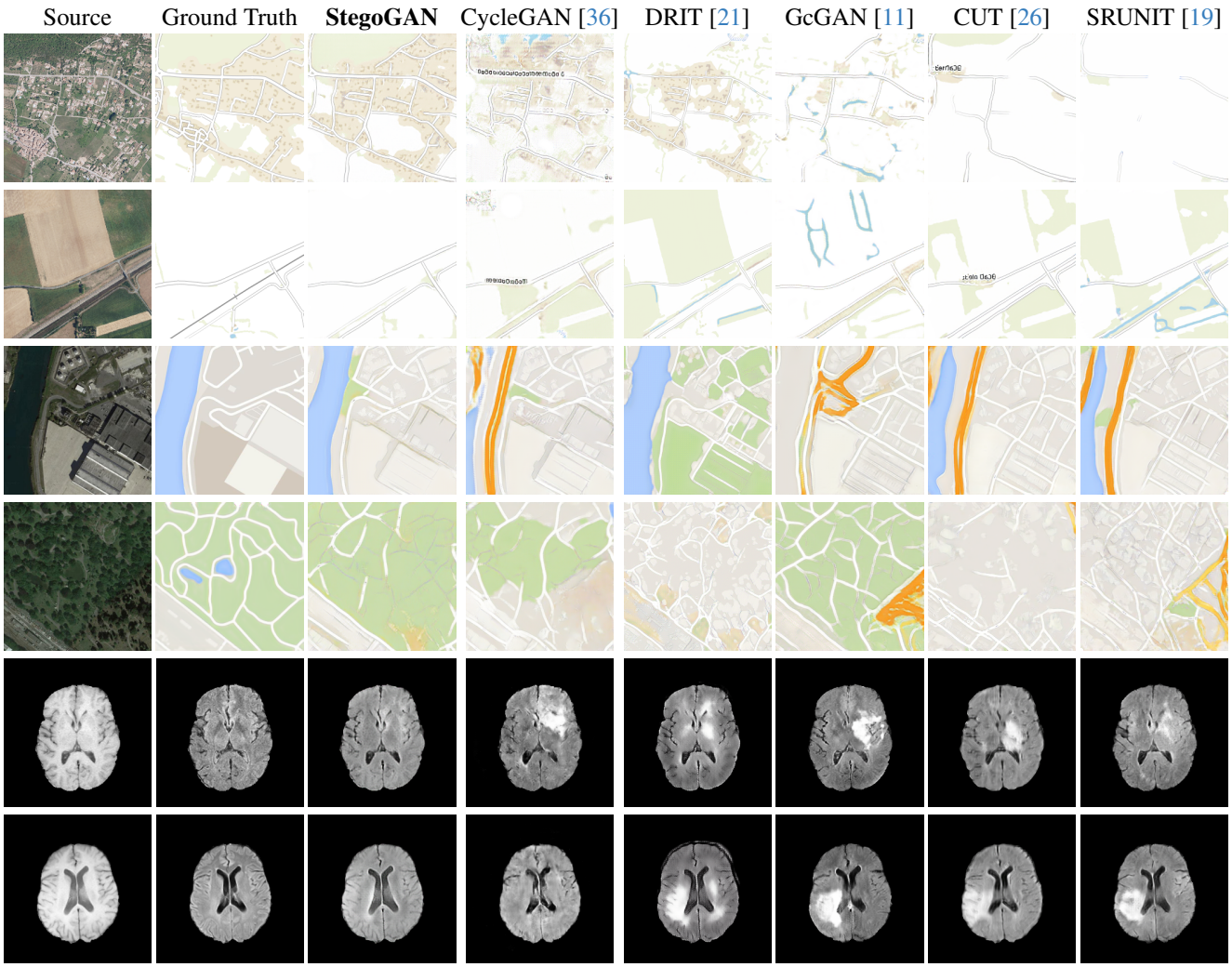


Figure 3. **Qualitative Comparison.** We report reconstructions from the test sets of PlanIGN (top two rows), GoogleMap (row 3 and 4), and MRI (last two rows). Contrary to the other models, StegoGAN does not hallucinate spurious toponyms, highways (orange roads), or tumors (white areas) and shows better semantic correspondences during translation.

value from $[0, 1]$ to $[1, 0]$, and upsamples the results to the dimensions of the input images.

$$I(m) = \text{upsample}\left(1 - \max_c m\right) . \tag{14}$$

The obtained consistency masks $I(m)$ have values close to 1 for pixels with only matchable content, and close to 0 otherwise. This enables us to define a loss for $y_{\text{gen}}^{\text{clean}}$ and $y_{\text{rec}}^{\text{clean}}$ that focuses solely on regions with matchable features:

$$\begin{aligned}\mathcal{L}_{\text{match}} = \mathbb{E}_{y\sim\mathcal{Y}} \| I\left(M(z_{\text{gen}})\right) \odot \left(y_{\text{gen}} - y_{\text{gen}}^{\text{clean}}\right) \| \\ + \mathbb{E}_{x\sim\mathcal{X}} \| I\left(M(z_{\text{rec}})\right) \odot \left(y_{\text{rec}} - y_{\text{rec}}^{\text{clean}}\right) \| .\end{aligned} \tag{15}$$

**Final Objective.** In addition to the standard CycleGAN loss components (4-5), we integrate $\mathcal{L}_{\text{match}}$ and $\mathcal{L}_{\text{reg}}$ into the overall loss function $\mathcal{L}(G_{\mathcal{X}\mapsto\mathcal{Y}}, G_{\mathcal{Y}\mapsto\mathcal{X}})$, weighted by their respective coefficients $\lambda_{\text{reg}}$ and $\lambda_{\text{match}}$. Crucially, our proposed approach remains unsupervised, requiring neither aligned images from $\mathcal{X}$ and $\mathcal{Y}$ nor specific annotations of unmatchable features.

## 4. Experiments

In this section, we assess the improvements brought by our method for non-bijective image translation across various datasets and compare them with existing models, both qualitatively and quantitatively.

**Implementation details.** We follow the setting of CycleGAN [36] as our baseline model: the generators are Resnets [15] and the discriminator is based on PatchGAN [18]. We define the encoders as the first half of the generator's layers and the decoders as the second half. The

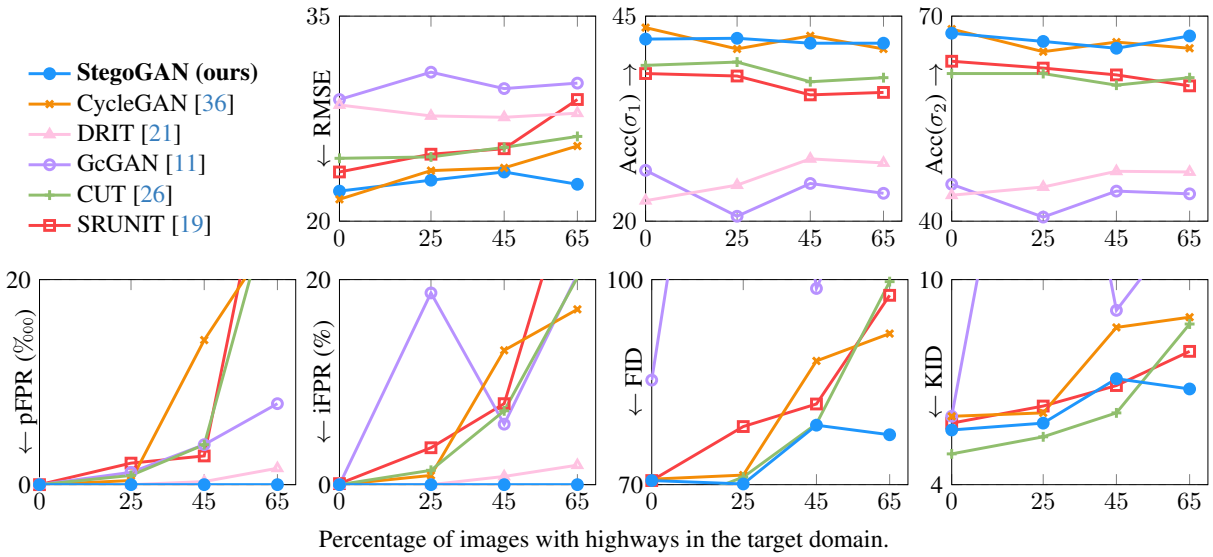


Figure 4. **Results on GoogleMaps.** We report the performance of several top-performing image translation models for different ratios of unmatchable features in the target domain of the training set. StegoGAN handles higher ratios better than competing methods.

unmatchability mask predictor $M$ is defined as a small 3-layer convolutional neural network (CNN). We set $\lambda_{\text{cyc}} = 10$, $\lambda_{\text{idt}} = 0.5$ as in [36] and $\lambda_{\text{match}} = 1$. The amplitude of the perturbation $\epsilon$ is 0.01 as in [7]. The Appendix provides more architecture and training details.

### 4.1. Datasets

We assess the performance of StegoGAN across several image translation tasks that feature unmatchable classes. Each dataset follows a consistent structure: the training set includes images from the source domain $\mathcal{X}$ devoid of a specific class (*e.g.*, tumors, motorways, toponyms) while the target domain $\mathcal{Y}$ does include that class. The test set comprises *paired* images from both domains, specifically excluding the unmatchable class. This setup allows us to quantify the models' hallucinations: any generated instances of the unmatchable class are necessarily spurious and due to its presence in the training set. We release all three curated datasets on the Zenodo platform and provide details below.

**PlanIGN.** $\mathcal{X}$ : **Aerial Photo,** $\mathcal{Y}$ : **Maps,** $k_{\mathcal{Y}}^{\text{unmatch}}$ : **Text.** We construct a dataset using open data from the French National Mapping Agency (IGN), comprising 1900 aerial (ortho-)images at 3m spatial resolution, and two versions of their corresponding topographic maps: one with and one without toponyms. This dataset presents a clear unmatchable class: place names. The training set includes 1000 maps with toponyms and 1000 aerial images, while the test set comprises 900 map samples without toponyms and their corresponding aerial images.

**GoogleMaps.** $\mathcal{X}$ : **Aerial Photo,** $\mathcal{Y}$ : **Maps,** $k_{\mathcal{Y}}^{\text{unmatch}}$ : **Highways.** The GoogleMaps dataset [18] is a standard benchmark for image translation tasks [19, 20]. It contains 1096 map/image pairs for training and 1098 for testing. To create a controlled non-bijective scenario, we exclude all satellite images that show highways and sample the maps of the training set to contain varying proportions of maps with highways, ranging from 0% to 65%, for a fixed total of 548 maps. For the test set we selected 898 pairs without highways.

**Brats MRI** $\mathcal{X}$ : **T1 Scans,** $\mathcal{Y}$ : **FLAIR,** $k_{\mathcal{Y}}^{\text{unmatch}}$ : **Tumors.** Lastly, we used a dataset of brain MRI scans [25] with two modalities: T1 (naive) and FLAIR (T2 Fluid Attenuated Inversion Recovery) [14]. We adapt the protocol that Cohen *et al*. [8] used for the Brats2013 datasets [24] to the more recent Brats2018 [2] dataset by varying the percentage of scans with tumors in the target domain. We selected transverse slices from the 60° to 100° range in the caudocranial direction [1] for both T1 and FLAIR scans. Each scan was classified as tumorous if more than 1% of its pixels were labeled as such, and as healthy if it contained no tumor pixels. The training set contains 800 images from each modality, with all source images (T1) being healthy and the target domain (FLAIR) comprising 60% tumorous scans. The test set contains 335 paired scans of healthy brains.

### 4.2. Evaluation metrics

We use a broad range of metrics to evaluate the performance of StegoGAN and other image translation algorithms in the non-bijective setting:

**FID and KID.** The Fréchet Inception Distance (FID) [16] and Kernel Inception Distance (KID) [4] are widely used to quantify the similarity between the distributions of real and generated images in the target domain.

Table 1. **Quantitative Comparison on PlanIGN**. Our model shows a remarkably better performance than other existing models.

| Method | RMSE↓ | Acc($\sigma_1$)↑ | Acc($\sigma_2$)↑ | FID↓ | KID↓ |
|---|---|---|---|---|---|
| CUT [26] | 30.5 | 46.7 | 55.8 | 68.4 | 2.8 |
| CycleGAN [36] | 27.0 | 15.1 | 57.6 | 97.5 | 6.6 |
| DRIT [21] | 34.8 | 33.6 | 36.9 | 76.4 | 3.8 |
| GcGAN [11] | 32.7 | 54.5 | 56.9 | 110.8 | 8.2 |
| SRUNIT [19] | 32.3 | 48.8 | 52.8 | 60.2 | **2.2** |
| **StegoGAN (ours)** | **22.5** | **66.1** | **74.8** | **58.4** | 2.4 |

**RMSE, Acc($\sigma_1$), and Acc($\sigma_2$).** As the test sets comprise paired images from both domains, we can directly compare the Root Mean Square Error (RMSE) between the real and predicted images in the target domain. We count a predicted pixel as correctly predicted if it deviate by less than a fixed threshold in any of the color channels [11]. We use $\sigma_1 = 5$ and $\sigma_2 = 10$ for the GoogleMap dataset and $\sigma_1 = 2$ and $\sigma_2 = 5$ for the less colorful PlanIGN dataset.

**pFPR and iFPR.** In the GoogleMap dataset, highways are always depicted in orange, allowing us to label pixels where all color channels differ by less than 20 units from $(240, 160, 30)$ as highways. In the Brats MRI dataset, we use a pretrained tumor detector [5] to find spurious tumors in the generated images. This allows us to compute the average false positive rate per pixel (pFPR) and per instance (iFPR) of the generated images.

### 4.3. Results

**Qualitative Results.** Figure 3 showcases StegoGAN's qualitative performance against other image translation algorithms. Notably, StegoGAN effectively avoids generating unmatchable classes such as texts, highways, and tumors, while producing high-quality image translations.

**Quantitative Results.** On the PlanIGN dataset (Table 1) and the Brats MRI dataset (Table 2), StegoGAN outperforms others in fidelity, achieving the lowest RMSE by a margin of 4.5 on PlanIGN and by 3.5 for Brats MRI. Furthermore, it significantly enhances pixel accuracy, with improvements of $+11.6$ in Acc($\sigma_1$) and $+17.2$ in Acc($\sigma_2$) on PlanIGN. In the MRI dataset, StegoGAN dramatically reduces false positive rates——over $20\times$ lower than CycleGAN and $10\times$ less than the next best model SRUNIT (for pFPR).

On the GoogleMap dataset, as shown in Figure 4, StegoGAN's performance is on par with CycleGAN at $0\%$ unmatchable cases and remains stable even as this ratio increases, unlike other methods that degrade. Remarkably, StegoGAN maintains a consistent false positive rate of $0$ across all tests, while this rate increases for all other methods.

**Unmatchability Masks.** In Figure 5, we illustrate the emergent ability of the unmatchability masks to trace the outline of unmatchable class instances like toponyms, highways, and tumors. This aspect highlights the versatility of our approach, which functions without explicit supervision or aligned images, offering a tool to explore the pairwise semantic differences between arbitrary datasets.

Table 2. **Quantitative Comparison on Brats MRI Flair → T1.** Our model outperforms competing method in terms of both reconstruction accuracy and consistency.

| Method | RMSE↓ | pFPR(‰)↓ | iFPR↓ | FID↓ | KID↓ |
|---|---|---|---|---|---|
| CUT [26] | 39.8 | 17.0 | 23.0 | 103.9 | 8.8 |
| CycleGAN [36] | 39.9 | 21.9 | 22.7 | 89.8 | 7.7 |
| DRIT [21] | 53.0 | 18.5 | 41.2 | 123.4 | 11.6 |
| GcGAN [11] | 41.7 | 24.7 | 22.4 | 61.9 | 3.5 |
| SRUNIT [19] | 42.5 | 15.1 | 21.8 | 58.9 | 2.9 |
| **StegoGAN (ours)** | **36.3** | **1.1** | **4.2** | **58.5** | **2.4** |

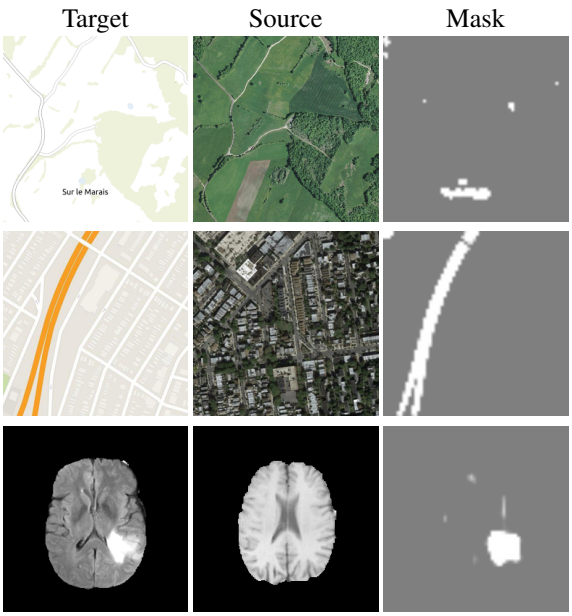


Figure 5. **Unmatchability Masks.** The unmatchability masks predicted in the backward cycle follow the instances of unmatchable features in the target domain: toponyms, highways, and tumors.

Table 3. **Ablation Study on Encoder Depth.** We evaluate the impact of changing the depth of the encoder on the reconstruction fidelity and the quality of unmatchability masks. Depth= -1 or 8 means no encoder or no decoder, respectively.

| Depth | Mask | | | Prediction | | | | |
|---|---|---|---|---|---|---|---|---|
| | mIOU↑ | Prec.↑ | recall↑ | RMSE↓ | A($\sigma_1$)↑ | A($\sigma_2$)↑ | FID↓ | KID↓ |
| -1 | 26.6 | 27.1 | **81.2** | **22.4** | 64.3 | 74.2 | 58.8 | 2.5 |
| 1 | 25.2 | 25.8 | **81.2** | 22.5 | **66.1** | **74.8** | **58.4** | **2.4** |
| 3 | 27.1 | 27.9 | 80.4 | 22.8 | 61.6 | 73.4 | 62.9 | 3.0 |
| 5 | 30.8 | 33.4 | 69.5 | 24.2 | 52.7 | 70.6 | 62.3 | 2.6 |
| 8 | **47.3** | **60.1** | 60.0 | 24.5 | 53.5 | 70.7 | 62.7 | 2.7 |

### 4.4. Ablation study and analysis

We explore the impact of our main design choices, as well as further capabilities and limitations of our approach. See the Appendix for further ablations.

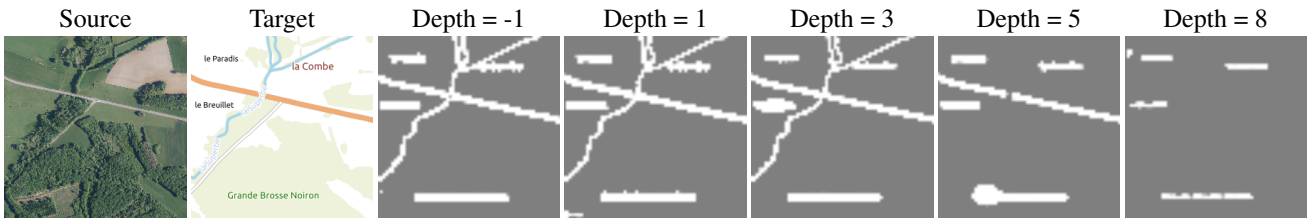


Figure 6. **Impact of Encoder Depth.** We visualize the unmatchability mask for encoders of different depths for the PlanIGN dataset. Shallower encoders consider more features as unmatchable.

**Encoder Ablation.** We conducted an ablation study on the definition of the intermediary representation $z$ by varying the depth at which the "encoder" ends and the "decoder" starts within the generator. Given that our generators consist of 9 consecutive convolutional blocks, we experimented with different configurations: $-1$ (indicating no decoder), 1 (the configuration used in our paper), as well as depths of 3, 5, and 8 (implying no decoder). We report in Table 3 the reconstruction error of these models, as well as the fidelity of the consistency mask with the toponym text mask. We observe that shallow encoders have better reconstruction accuracy while the consistency masks of deeper encoders better approximate the text masks.

Visualizing these masks in Figure 6, we observe that shallow encoders consider complex features such as highways and rivers as unmatchable features, while deeper encoders do not. Shallower encoders seem more influenced by the variation in appearance (*e.g.*, rivers being sometimes covered in vegetation or with varying colors) while deeper encoders focus on high-level semantics. We argue that both definitions are equally valid, and that varying the depth of the encoders can provide insights into the nature of the semantic mismatch between datasets.

**Parameterization.** In Table 4, we analyze the effects of omitting the additional terms $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{match}}$ from the loss function $\mathcal{L}$ in Equation (4). We show that while $\mathcal{L}_{\text{match}}$ generally yields modest improvements across all metrics, $\mathcal{L}_{\text{reg}}$ is pivotal, particularly for learning the target distribution. This outcome aligns with our expectations, as the absence of $\mathcal{L}_{\text{reg}}$ allows the network to transmit all information, matchable or not, to the $G_{\mathcal{Y}\mapsto\mathcal{X}}$ decoder without repercussions, impeding the training of the $G_{\mathcal{X}\mapsto\mathcal{Y}}$ encoder.

**Limitations.** We augment the CycleGAN framework with a module $M$ and two hyper-parameters $\lambda_{\text{match}}$ and $\lambda_{\text{reg}}$, thereby adding to the complexity of its training dynamics. Moreover, the concept of unmatchability, integral to our approach, is inherently subjective. Given enough semantic detail, any two distinct datasets could be considered unmatchable. As a result, fine-tuning the hyperparameter $\lambda_{\text{reg}}$ is essential to balance the elimination of hallucinations against the retention of necessary details: increasing its value leads to more conservative masks and improves the visual appearance of the generated images, at the cost of more spurious features. Visualizing the consistency mask is often a useful form of guidance. More details on learning strategies can be found in the Appendix.

Table 4. **Impact of Additional Loss Terms.** We evaluate on the GoogleMap dataset the effect of removing our proposed losses. $\mathcal{L}_{\text{match}}$ has a small impact while $\mathcal{L}_{\text{reg}}$ is pivotal.

| Settings | | RMSE↓ | Acc($\sigma_1$)↑ | Acc($\sigma_2$)↑ | FID↓ | KID↓ |
|---|---|---|---|---|---|---|
| $\mathcal{L}_{\text{reg}}$ | $\mathcal{L}_{\text{match}}$ | | | | | |
| ✓ | ✓ | 22.7 | 41.7 | 67.1 | 77.3 | 6.8 |
| ✓ | ✗ | 24.1 | 41.5 | 64.7 | 88.0 | 7.5 |
| ✗ | ✓ | 26.7 | 26.7 | 58.9 | 271.1 | 26.6 |
| ✗ | ✗ | 25.0 | 25.0 | 61.0 | 303.6 | 33.4 |

We also acknowledge the potential of recent denoising diffusion models for image-to-image translation tasks [22, 32]. While our method is not confined to GANs and could be adapted to diffusion models with cycle consistency losses [28, 31], unpaired image translation with diffusion models is a nascent field with unique challenges. We plan to explore this area in future research.

## 5. Conclusions

We have introduced StegoGAN, a model built upon the CycleGAN framework, which leverages the mechanism of steganography to address the challenges of non-bijective image-to-image translation. Our model demonstrates an improved capability to handle divergent distributions between domains, as evidenced by its performance across various datasets, including aerial imagery, topographic maps, and MRI scans. We hope that our work will inspire further research in the little-studied area of non-bijective image translation. We find this research direction inportant to ensure image translation models are transferable and applicable in real-world scenarios, where datasets rarely conform to the level of curation typically found in research benchmarks.

## References


[1] Simon Andermatt, Antal Horváth, Simon Pezold, and Philippe Cattin. Pathology segmentation using distributional differences to images of healthy origin. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part I 4*, pages 228–238. Springer, 2019. 6

[2] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018. 6

[3] Christoph Baur, Robert Graf, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Steganomaly: Inhibiting cyclegan steganography for unsupervised anomaly detection in brain mri. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–727. Springer, 2020. 2

[4] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018. 6

[5] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in Biology and Medicine*, 109, 2019. 7

[6] Sidonie Christophe, Samuel Mermet, Morgan Laurent, and Guillaume Touya. Neural map style transfer exploration with gans. *International Journal of Cartography*, 2022. 2

[7] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017. 2, 4, 6

[8] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 529–536. Springer, 2018. 1, 2, 6

[9] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022. 1, 2

[10] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of jpg compression on adversarial images. *ArXiv*, abs/1608.00853, 2016. 2

[11] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5, 6, 7

[12] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2, 3

[14] Joseph V Hajnal, David J Bryant, Larry Kasuboski, Pradip M Pattany, Beatrice De Coene, Paul D Lewis, Jacqueline M Pennock, Angela Oatridge, Ian R Young, and Graeme M Bydder. Use of fluid attenuated inversion recovery (FLAIR) pulse sequences in MRI of the brain. *Journal of computer assisted tomography*, 1992. 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[17] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1125–1134, 2017. 2, 5, 6

[19] Zhiwei Jia, Bodi Yuan, Kangkang Wang, Hong Wu, David Clifford, Zhiqiang Yuan, and Hao Su. Semantically robust unpaired image translation for data with unmatched semantics statistics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14273–14283, 2021. 2, 5, 6, 7

[20] Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18260–18269, 2022. 2, 6

[21] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018. 2, 5, 6, 7

[22] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *CVPR*, 2023. 8

[23] Yu Li, Sheng Tang, Rui Zhang, Yongdong Zhang, Jintao Li, and Shuicheng Yan. Asymmetric gan for unpaired image-to-image translation. *IEEE Transactions on Image Processing*, 28(12):5881–5896, 2019. 2

[24] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 6

[25] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014. 6

[26] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020. 2, 5, 6, 7

[27] Luca Sestini, Benoit Rosa, Elena De Momi, Giancarlo Ferrigno, and Nicolas Padoy. Fun-sis: A fully unsupervised approach for surgical instrument segmentation. *Medical Image Analysis*, 85:102751, 2023. 2

[28] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *ICLR*, 2023. 8

[29] Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. AttentionGAN: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE transactions on neural networks and learning systems*, 2021. 2

[30] Weilun Wang, Wen gang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14000–14009, 2021. 2

[31] Chen Henry Wu and Fernando De la Torre. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *ICCV*, 2023. 8

[32] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, Radu Timotfe, and Luc Van Gool. Diffi2i: Efficient diffusion model for image-to-image translation. *arXiv preprint arXiv:2308.13767*, 2023. 8

[33] Shaoan Xie, Qirong Ho, and Kun Zhang. Unsupervised image-to-image translation with density changing regularization. *Advances in Neural Information Processing Systems*, 35:28545–28558, 2022. 2

[34] Zongben Xu, Hai Zhang, Yao Wang, XiangYu Chang, and Yong Liang. L 1/2 regularization. *Science China Information Sciences*, 53:1159–1169, 2010. 4

[35] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. Unpaired portrait drawing generation via asymmetric cycle mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '20)*, pages 8214–8222, 2020. 2

[36] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017. 1, 2, 3, 5, 6, 7