mathematics

MDPI

Article

Probability Density Function Distance-Based Augmented CycleGAN for Image Domain Translation with Asymmetric Sample Size

Lidija Krstanović [1], Branislav Popović [2], Sebastian Baloš [3], Milan Narandžić [2] and Branko Brkljač [2,*]

[1] Department of Fundamental Disciplines in Engineering, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia; lidijakrstanovic@uns.ac.rs

[2] Department of Power, Electronic and Telecommunication Engineering, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia; bpopovic@uns.ac.rs (B.P.); orange@uns.ac.rs (M.N.)

[3] Department of Production Engineering, Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia; sebab@uns.ac.rs

* Correspondence: brkljacb@uns.ac.rs

Abstract: Many image-to-image translation tasks face an inherent problem of asymmetry in the domains, meaning that one of the domains is scarce—i.e., it contains significantly less available training data in comparison to the other domain. There are only a few methods proposed in the literature that tackle the problem of training a CycleGAN in such an environment. In this paper, we propose a novel method that utilizes pdf (probability density function) distance-based augmentation of the discriminator network corresponding to the scarce domain. Namely, the method involves adding examples translated from the non-scarce domain into the pool of the discriminator corresponding to the scarce domain, but only those examples for which the assumed Gaussian pdf in VGG19 net feature space is sufficiently close to the GMM pdf that represents the relevant initial pool in the same feature space. In experiments on several datasets, the proposed method showed significantly improved characteristics in comparison with a standard unsupervised CycleGAN, as well as with Bootstraped SSL CycleGAN, where translated examples are added to the pool of the discriminator corresponding to the scarce domain, without any discrimination. Moreover, in the considered scarce scenarios, it also shows competitive results in comparison to fully supervised image-to-image translation based on the pix2pix method.

Keywords: CycleGAN; domain translation; selective data augmentation; bootstrapping; pdf distance

MSC: 68T20; 68T05; 68T07; 68U10; 94A08

check for updates

Academic Editors: Volodymyr Ponomaryov, Vladimir Lukin, Bogdan Smolka and Beatriz P. García Salgado

Received: 22 February 2025
Revised: 3 April 2025
Accepted: 14 April 2025
Published: 25 April 2025

Citation: Krstanović, L.; Popović, B.; Baloš, S.; Narandžić, M.; Brkljač, B. Probability Density Function Distance-Based Augmented CycleGAN for Image Domain Translation with Asymmetric Sample Size. *Mathematics* **2025**, *13*, 1406. https://doi.org/10.3390/math13091406

1. Introduction

Image-to-image (I2I) translation tasks are a key component of many image processing, computer graphics, and computer vision problems, as well as other similar problems. Some I2I methods that have been proposed are given in, for example, [1–4] for semantic image synthesis, [5–8] for image-to-image translation, and [9,10] for image super-resolution. They consist of constructing the mapping that translates images from one (source) domain to another (target) domain (or many of these), thus preserving the content of the image while the style of the image belonging to the first domain is changed to that of the second domain. The best performances of the I2I translations are obtained in the fully supervised training scenario, where in the training phase, all images in the source and the target (scarce) domain are assumed to be available in pairs (paired images). Those methods were developed first (see [11] for the pix2pix learning strategy) and are mostly based on conditional Generative Adversarial Networks (cGANs). However, the method was further developed by [12–14], whose solutions still failed to capture the complex structural relationships of scenes in cases when two domains had drastically different views while trying to achieve mapping through a single translation network. Moreover, despite improvements, the main drawback of supervised methods is related to the insufficient amount of paired image data that are available in most real-world I2I translation problems and the high cost of creating such datasets.

In order to cope with the previously mentioned problem, an unsupervised CycleGAN I2I method was proposed in [15]. It uses two GANs oriented in opposite directions, i.e., from one domain into another, and vice versa. The problem of highly under-constrained mappings, which are introduced in this approach, was dealt with by introducing a cycle consistency loss, which forces the mentioned mappings to be as close to bijective as possible. The method was proved to be very effective for preserving the semantic information of the data with respect to I2I translations, as well as other domain transfer tasks, such as the following: image-to-image translation [15], emotion style transfer [16], and speech enhancement [17]. Nevertheless, all of those mentioned, as well as many other domain transfer tasks, have inherent domain asymmetry, meaning that one of the domains has significantly less available training data (noted as scarce domain). There have been few studies reported in the literature devoted to resolving this problem. In [18], an augmented cyclic adversarial learning model that enforces the cycle consistency constraint via an external task-specific model is proposed. Additionally, in [19], the authors add semi-supervised task-specific attention modules to generate images that are used for improving performance in a medical image classification task. Recently, in [20], a bootstrapped SSL CycleGAN architecture was proposed, where the aforementioned problem was overcome using the following two strategies. Firstly, by using a relatively small percentage of available labeled training data from the reduced (scarce) domain and a Semi-Supervised Learning (SSL) approach, the method prevents overfitting of the discriminator belonging to the reduced domain, which would otherwise occur during initial training iterations due to the small amount of available training data in the scarce domain. Secondly, after initial learning guided by the described SSL strategy, additional augmentation of the reduced data domain is performed by inserting artificially generated training examples into the training poll of the data discriminator belonging to the scarce domain. Bootstrapped samples are generated by the neural network that performs the transfer from the fully observable domain to the scarce domain with its currently trained parameters. Moreover, in [21], the method for image translation is adapted for an application in which the fully observable image domain contains additional semantic information (in the form of associated text inputs). This presence of cross-modal information makes possible a different learning strategy design, in comparison to our method, which was designed and extensively tested for exclusively visual inputs. In [22], the fully observable domain has clearly distinguishable object classes, which correspond to similar categories, enabling fine-grained partitioning of the image domain into disjoint subsets (modes) for each of the known classes.

The problem of “imbalanced” or asymmetric sample size is also present in the tasks of training a multi-class classifier with an unequal number of training instances per category, e.g., in [23–25], where the CycleGAN model is used to compensate for an unequal number of training instances per category in different classification tasks that are of interest. However, these types of problems differ from the one investigated in this study.

In this work, we propose an extension of the methodology given in [20]. Namely, instead of just adding all of examples that are translated from the fully observable to the scarce domain (translated images of the training samples from the fully observable domain that do not have a pair in the scarce domain) and adding them to the training poll of the data discriminator belonging to the scarce domain, as implemented in [20], we implement this process selectively, in a more subtle manner. We actually add only those translated examples whose probability density function (pdf) in some predefined feature space is sufficiently close, in terms of distances or similarity measures between pdfs of feature vectors, to the pdf that represents the original pool of discriminator data in the scarce domain, obtained in the same feature space. To achieve an adequate feature space, we utilize a pre-trained VGG19 convolutional neural network (CNN) and extract feature maps from certain network layers [26].

The primary concept of the method proposed in this paper is to expand the pool of discriminator training samples in the scarce domain by adding translated images of unpaired samples from the fully observable domain. However, only images translated from the fully observable domain that have a pdf for their feature vectors (CNN-based image representations) that closely resembles (in terms of pdf similarity measure) the pdf of training samples that have already been assigned to the discriminator in the scarce domain (within the same CNN based feature space) will be included.

For the pdf representing the translated example that has the potential to be added to the pool of the discriminator of the scarce domain, we assume a multivariate Gaussian distribution and estimate its mean and covariance using the maximum likelihood (ML) method. On the other hand, for the pdf that represents the actual data from the discriminator pool (original images from the scarce domain), we assume a Gaussian mixture model (GMM) of their CNN-based representations with some small number of components, which is estimated over the same feature space obtained by the convolutional layers of the VGG19 CNN. Although several GMM similarity measures can be used to compare GMM pdfs, in this paper, we use one of the computationally most efficient [27–29]. The choice of feature space is also important. For this purpose, we chose the reshaped tensors representing convolutional feature maps from the layers of the pretrained VGG19 CNN proposed in [26] (e.g., see also [30,31]) as image features. The efficiency of this approach has already been demonstrated in various image recognition tasks, as well as image style transfer tasks (see [32–34]). Similar to what was done in [20], we use semi-supervised learning (SSL) on a predefined amount of available paired data to ensure that the generator performing translation to the scarce domain becomes sufficiently well trained. The initial SSL strategy that is based on a small set of paired observations allows the discriminator as well as the generator that are residing in the scarce domain to avoid overfitting and learn necessary parameters to some extent. After a sufficient number of iterations, the corresponding generator of the fully observable domain of the CycleGAN is periodically called to translate additional examples to the scarce domain (first phase of the propose bootstrapping process), from which some of them are chosen to be added in the training pool of the discriminator of the scarce domain (second phase), based on the previously mentioned pdf distance criteria.

This paper is organized as follows: In Section 2, we give a brief description of the baseline CycleGAN, first proposed in [15], as well as the bootstrapped SSL (BTS-SSL) CycleGAN proposed in [20]. In Section 3, the novel pdf distance-based augmented CycleGAN for asymmetric image domains (*PdfAugCycleGAN* further in the text) is proposed and described. In Section 4, the experimental results and comparisons of the proposed *PdfAugCycleGAN* to the baseline *CycleGAN* and *BTS-SSLCycleGAN*, as well as the fully supervised *pix2pix* method [11], are presented on several real datasets with varying amounts of data in the scarce domain. Finally, in Section 5, we provide the corresponding conclusions. The list of mathematical symbols is provided in Appendix A.

2. Baseline CycleGAN Methods

In [35], a non-parametric method called a Generative Adversarial Network (GAN) that learns the true data distribution based on competitive learning of two networks (generator and discriminator) was presented and has since quickly become ground-breaking in many applications of machine learning involving generative models. The key ingredient that underlies the game theoretic nature of the method is the Nash equilibrium, expressed by the minimax loss of the training procedure given by the following cross-entropy type loss:

$$\min_G \max_D \mathbb{E}_{y \sim p_Y(y)} \ln(D(y)) + \mathbb{E}_{z \sim p_Z(z)} \ln(1 - D(G(z))) \tag{1}$$

Namely, adversarial training consists of the discriminator network D trying to discriminate between the synthetic samples artificially generated by the generator network G and the ground truth observations y available in the training data. Generator G is trying to deceive the discriminator by providing synthetic examples that are as similar as possible to the ground truth.

To perform I2I translation (and other domain transfer tasks), a supervised (using paired examples) version of the GAN architecture, named conditional GAN (cGAN), was proposed in [36] and applied to I2I in [11], named the *pix2pix* method. Namely, in contrast, conditional GANs learn a mapping from an observed image x and a random noise vector z to y, i.e., $G : (x, z) \to y$. Adversarial training is similar to the case of GAN, except that both the discriminator D and the generator G have input (x, z), contrary to the case of an ordinary GAN network:

$$\min_G \max_D \mathbb{E}_{x \sim p_X(x), y \sim p_Y(y)} \ln(D(x, y)) + \mathbb{E}_{x \sim p_X(x), z \sim p_Z(z)} \ln(1 - D(x, G(x, z))), \tag{2}$$

where p_X and p_Y are data distributions that correspond to domains X and Y, respectively, while p_Z is a distribution of the latent variable z (commonly a normal distribution).

Although the *pix2pix* method showed superior performance in the presence of a large number of paired examples, in real applications, it is very hard to obtain such a large amount of labeled data, as this requires significant annotation effort, which is hard to obtain. The CycleGAN network emerged in [15] in order to perform I2I, as well as style transfer from one domain into another, without using any supervisor, i.e., paired data examples. This is done by invoking the *cycle consistency* loss in the overall loss function by designing two domain translators or mappers $G : X \to Y$ and $F : Y \to X$ in mutually opposite directions, where X and Y denote two image domains, i.e., style transfer domains. Cycle consistency loss encourages mappings F and G to be as close to a bijection (F and G are inverse to each other) to a sufficient extent, i.e., by making $G(F(y)) \approx y$, as well as $F(G(x)) \approx x$. It defines the diameter of the regions in X that are mapped by G to the same point $y \in Y$, and in principle, the same for F. If we denote the discriminator networks corresponding to domains X and Y by D_X and D_Y, respectively, we utilize the following adversarial costs:

$$\begin{aligned} \mathcal{L}_{adv}(G, D_Y) &= \mathbb{E}_{y \sim p_Y(y)}[\ln(D_Y(y))] + \mathbb{E}_{x \sim p_X(x)}(1 - D_Y(G(x))) \\ \mathcal{L}_{adv}(F, D_X) &= \mathbb{E}_{x \sim p_X(x)}[\ln(D_X(x))] + \mathbb{E}_{y \sim p_Y(y)}(1 - D_X(F(y))), \end{aligned} \tag{3}$$

as well as the cycle consistency cost:

$$\mathcal{L}_{cycle}(G, F) = \mathbb{E}_{x \sim p_X(x)} \|F(G(x)) - x\|_{l_1} + \mathbb{E}_{y \sim p_Y(y)} \|G(F(y)) - y\|_{l_1}, \tag{4}$$

so that the full optimization cost optimized by the learning strategy of CycleGAN as follows:

$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{adv}(G, D_Y) + \mathcal{L}_{adv}(F, D_X) + \lambda_{cycle}\mathcal{L}_{cycle}(G, F), \tag{5}$$

where $\lambda_{cycle} > 0$ controls the forcing of the cycle consistency.

The basic CycleGAN training procedure is given by the following pseudo-code in Algorithm 1:

Algorithm 1 CycleGAN training procedure

```
procedure CycleGAN
    $N$—number of iterations; $m$—minibatch size; $\eta > 0$, learning rate; $X \in \mathcal{X}$, $Y \in \mathcal{Y}$, unpaired or unlabeled training sets, such that $|X| \ll |Y|$;
    Randomly initialize the parameters of the discriminators $D_X$, $D_Y$, and generators $G_{X\to Y}$, $G_{Y\to X}$: $\theta_{D_X}, \theta_{D_Y}, \theta_{G_{X\to Y}}, \theta_{G_{Y\to X}}$
    for $k = 1$ to $N$ do
        Sample minibatch of unpaired training data $\{x_1, \dots, x_m\} \subset X$, $\{y_1, \dots, y_m\} \subset Y$
        $\widehat{\mathcal{L}}_{adv}(G_{X\to Y}, D_Y) = \frac{1}{m}\sum_{i=1}^{m} \ln D_Y(y_i) + \frac{1}{m}\sum_{i=1}^{m} \ln\left(1 - D_Y(G_{X\to Y}(x_i))\right)$
        $\widehat{\mathcal{L}}_{adv}(G_{Y\to X}, D_X) = \frac{1}{m}\sum_{i=1}^{m} \ln D_X(x_i) + \frac{1}{m}\sum_{i=1}^{m} \ln\left(1 - D_X(G_{Y\to X}(y_i))\right)$
        $\theta_{D_X}^{(k+1)} \longleftarrow \theta_{D_X}^{(k)} - \eta\, \nabla_{\theta_{D_X}} \widehat{\mathcal{L}}_{adv}(G_{Y\to X}, D_X)$
        $\theta_{D_Y}^{(k+1)} \longleftarrow \theta_{D_Y}^{(k)} - \eta\, \nabla_{\theta_{D_Y}} \widehat{\mathcal{L}}_{adv}(G_{X\to Y}, D_Y)$
        $\theta_{G_{Y\to X}}^{(k+1)} \longleftarrow \theta_{G_{Y\to X}}^{(k)} - \eta\, \nabla_{\theta_{G_{Y\to X}}} \left[\widehat{\mathcal{L}}_{adv}(G_{Y\to X}, D_X) + \lambda_{cyc}\, \mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X})\right]$
        $\theta_{G_{X\to Y}}^{(k+1)} \longleftarrow \theta_{G_{X\to Y}}^{(k)} - \eta\, \nabla_{\theta_{G_{X\to Y}}} \left[\widehat{\mathcal{L}}_{adv}(G_{X\to Y}, D_Y) + \lambda_{cyc}\, \mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X})\right]$
    end for
end procedure
```

In the semi-supervised learning (SSL) case, i.e., in the case where there is a limited amount of paired (labeled) data, the additional SSL cost is added to (5), defined as follows:

$$\mathcal{L}_{SSL}(G, F) = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} [\|G(x_p) - y_p\|_{l_1} + \|F(y_p) - x_p\|_{l_1}], \tag{6}$$

with $\mathcal{P}$ denoting set of indices of paired data samples. Thus, the overall cost of SSL CycleGAN is given as follows:

$$\begin{aligned} \mathcal{L}(G, F, D_X, D_Y) &= \mathcal{L}_{adv}(F, D_X) + \mathcal{L}_{adv}(G, D_Y) \\ &+ \lambda_{cycle}\mathcal{L}_{cycle}(G, F) + \lambda_{SSL}\mathcal{L}_{SSL}(G, F), \end{aligned} \tag{7}$$

with $\lambda_{SSL} > 0$ controlling the influence of SSL cost on the overall cost.

Bootstrapped SSL CycleGAN for Asymmetric Domain Transfer

In order to deal with the mutually asymmetric domains, i.e., a problem where one of the domains involved in the translation task is scarce, meaning that it lacks a sufficient amount of data, Bootstrapped SSL CycleGAN (*BTS-SSLCycleGAN*) is proposed in [20]. It utilizes two concepts in addition to the standard *CycleGAN*. The first is SSL training, which uses cost (7) and assumes that some number of paired training samples is preventing overfitting of the discriminator network during the initial iterations of the learning process. The second concept is a bootstrapping strategy that aims to overcome the difference in the amount of training data between the scarce and non-scarce domains by artificially expanding the amount of the unlabeled training pool of the discriminator D_X on the scarce domain X. This is achieved as follows.

In the initial phase of the training procedure, the parameters of the *CycleGAN* model are first optimized for some time using the previously mentioned SSL strategy. After the initial training of the generator $G_{Y\to X}$ (i.e., F), and when it is considered "reliable enough", it is used as a bootstrapping sampler for data augmentation of the discriminator D_X. This is done by periodically translating a predefined percentage (sampled by uniform distribution) of available examples from the domain Y to the scarce domain X and adding those to the pool of the discriminator D_X, thus bootstrapping the statistics $\hat{p}_X(x)$ of the scarce domain X, which approximates the ground truth $p_X(x)$.

3. Pdf Distance-Based Augmented CycleGAN

In this section, we propose and describe a novel method we call pdf distance-based augmented CycleGAN (*PdfDistCycleGAN*). In this approach, we expand on the previous idea of *BTS-SSLCycleGAN* by adding to the data pool of the discriminator D_X of the scarce domain X only those translated samples from the domain Y which agree in some way, which we will specify soon in the text, and to some extent, to the currently estimated pdf $\hat{p}_X(x)$ of the scarce domain X. We do this implicitly by measuring the similarity between the Gaussian pdf that corresponds to the particular bootstrapping sample translated from Y to X to the GMM that corresponds to the pool of the discriminator D_X in some specified feature space. We specify the previous information as follows.

3.1. Proposed PdfAugCycleGAN Data Augmentation Process

We utilize the feature maps of the pretrained VGG19 convolutional network *VGG19 Net*, which is 19 layers deep and trained on the ImageNet database [30,31], to obtain the feature space in which pdf similarity is measured, i.e., the actual domain of p_X. For the translated example $\hat{x} = F(y)$, for some particular $y \in Y$, we bring $\hat{x}$ to the input of the instance of a pre-trained VGG19 CNN, and we use the obtained feature map tensor $T^{(l),\hat{x}} \in \mathbb{R}^{m^{(l)}\times n^{(l)}\times d}$ from some specified l-th feature map level (or more of those). Thus, by vectorizing $T^{(l),\hat{x}}$, we obtain the set of d dimensional feature vectors $T^{(l),\hat{x}}_{vct} = \{f^{(l),\hat{x}}_{11}|\ldots|f^{(l),\hat{x}}_{m^{(l)}n^{(l)}}\}$ from which, by assuming that these are generated by single multivariate Gaussian pdf $\mathcal{N}(\mu^{(l),\hat{x}}, \Sigma^{(l),\hat{x}})$, by using the ML technique [37,38], we obtain the estimates $(\hat{\mu}^{(l),\hat{x}}, \hat{\Sigma}^{(l),\hat{x}})$ of these as follows:

$$\hat{\mu}^{(l),\hat{x}} = \frac{1}{m^{(l)}n^{(l)}}\sum_{i=1}^{m^{(l)}}\sum_{j=1}^{n^{(l)}} f^{(l),\hat{x}}_{ij}$$
$$\hat{\Sigma}^{(l),\hat{x}} = \frac{1}{m^{(l)}n^{(l)}-1}\sum_{i=1}^{m^{(l)}}\sum_{j=1}^{n^{(l)}}\left(f^{(l),\hat{x}}_{ij} - \mu^{(l),\hat{x}}\right)\left(f^{(l),\hat{x}}_{ij} - \mu^{(l),\hat{x}}\right)^T \tag{8}$$

On the other hand, for the pool of existing training samples of the discriminator D_X corresponding to the scarce domain X, we assign the Gaussian mixture pdf:

$$f_{D_X} = \sum_{i=1}^{M}\alpha_i\mathcal{N}(\mu_{x_i}, \Sigma_{x_i}) + \sum_{j=1}^{M_{add}}\beta_j\mathcal{N}(\mu_{\hat{x}^{add}_j}, \Sigma_{\hat{x}^{add}_j}),$$
$$\alpha_i, \beta_j \geq 0,\ \sum_{i=1}^{M}\alpha_i + \sum_{j=1}^{M_{add}}\beta_j = 1,\ \text{for all } i, j \tag{9}$$

where M is the number of original image examples x_i of the pool corresponding to D_X, while M_{add} is the number of translated examples $\hat{x}^{add}_j = G(y_j)$, which have already been added to the pool corresponding to D_X. Estimates $\mu_{\hat{x}^{add}_j}$ and $\Sigma_{\hat{x}^{add}_j}$, corresponding to previously translated and added bootstrapping image samples $\hat{x}_j$, are obtained using VGG19 features, as described by (8). The same also holds for original image samples x_i in (9).

In this way, all parameters of single multivariate Gaussian distributions that make GMM are estimated exclusively from feature vectors of individual image samples. Thus, on the domain level, Gaussians corresponding to individual image samples are combined into a Gaussian mixture.

The actual mechanics of obtaining the component weights in GMM are given as follows. We assign $\tilde{\alpha}_i = \alpha$, $i = 1 \ldots, M$, $\tilde{\beta}_j = \beta$, $j = 1, \ldots, M_{add}$, $\alpha > \beta > 0$, with:

$$\alpha_i = \frac{\tilde{\alpha}_i}{\sum_{i=1}^{M} \tilde{\alpha}_i + \sum_{j=1}^{M_{add}} \tilde{\beta}_j}, \; i = 1 \ldots, M, \tag{10}$$

$$\beta_j = \frac{\tilde{\beta}_j}{\sum_{i=1}^{M} \tilde{\alpha}_i + \sum_{j=1}^{M_{add}} \tilde{\beta}_j}, \; j = 1, \ldots, M_{add}, \tag{11}$$

so that the constraints from the Equation (9) in the original manuscript still hold. Thus, we assign higher weights to the original examples belonging to the scarce domain, as we consider those to be more significant than the ones that are translated. We use $\beta = \frac{1}{2}\alpha$.

Another, more sophisticated approach is to weight individual contributions of the translated examples according to the distance of their corresponding Gaussians to the currently estimated GMM f_{D_X} of the training samples that are already in the training pool of the discriminator in the scarce domain, which is given by Equation (9) in the original manuscript. Of course, in this case, we assign equal weights for all examples that are already in the training pool of the discriminator in the scarce domain X. One possible solution is given by the following weighting scheme:

$$\alpha_i = \frac{1}{2M}, \tag{12}$$

$$\beta_j = \frac{1}{2} \frac{d_{sim}(g_j, f_{D_X})}{\sum_{\tilde{j}=1}^{M_{add}} d_{sim}(g_{\tilde{j}}, f_{D_X})}, \tag{13}$$

where $d_{sim}(\cdot, \cdot)$ denotes utilized GMM similarity measure (see Section 3.3), and g_j is the Gaussian pdf obtained in VGG19 feature space, corresponding to the particular translated example $\hat{x}_j \in X$, where it holds that $\hat{x}_j = F(y_j)$, $y_j \in Y$. Thus, Equation (9) from the original manuscript still holds, but with more appropriate weight assignment. Nevertheless, in the experiments that we have conducted, we obtained unnoticeable differences in accuracy on the tested datasets. Therefore, we have presented only results with the weighting methodology presented in (10) and (11).

To decide whether or not the translated example $\hat{x}_j$ is to be added to the pool of discriminator D_X, Figure 1, some particular GMM similarity measure $d_{sim}(g_j, f_{D_X})$ between the Gaussian $\mathcal{N}(\mu_{\hat{x}_j}, \Sigma_{\hat{x}_j})$, corresponding to the translated example $\hat{x}_j$ (mixture is then defined as $g_j = 1 \cdot \mathcal{N}(\mu_{\hat{x}_j}, \Sigma_{\hat{x}_j})$), and the mixture f_{D_X}, representing the pool of the discriminator D_X, is evaluated, where both mixtures are over the same specified CNN feature space (e.g., layer l values after forward pass in VGG19).

The adopted bootstrapping strategy (BTS) is that some predefined percentage P of the lowest score $d_{sim}(g_j, f_{D_X})$ translated examples, obtained via translation from unpaired examples y_j belonging to the non-scarce domain Y, are added to the pool of discriminator D_X periodically. We note that the parameters of the CNN implementing translation mapping F are the currently trained parameters, i.e., the parameters trained up to the moment when the translation of the novel examples into the scarce domain occurs.

Thus, the pool of discriminator D_X is periodically boosted based on newly added images generated via image translation from the unpaired examples in the fully observable domain. The translation mapping is performed based on the currently determined parameters of the translation network, but only those translated examples that do not "spoil" the original distribution p_X are added to the pool corresponding to D_X, thus preventing the presence of outliers in the augmentation process. Next, the generator $F : Y \to X$, as well as the discriminator D_X are trained on the augmented pool of the discriminator D_X by using the adversarial cost $\mathcal{L}_{adv}(F, D_X)$ from (3), as well as the cycle consistency cost (4) and SSL cost (7). This update is also made periodically, after the previously described process of data augmentation of the scarce domain X is finished.

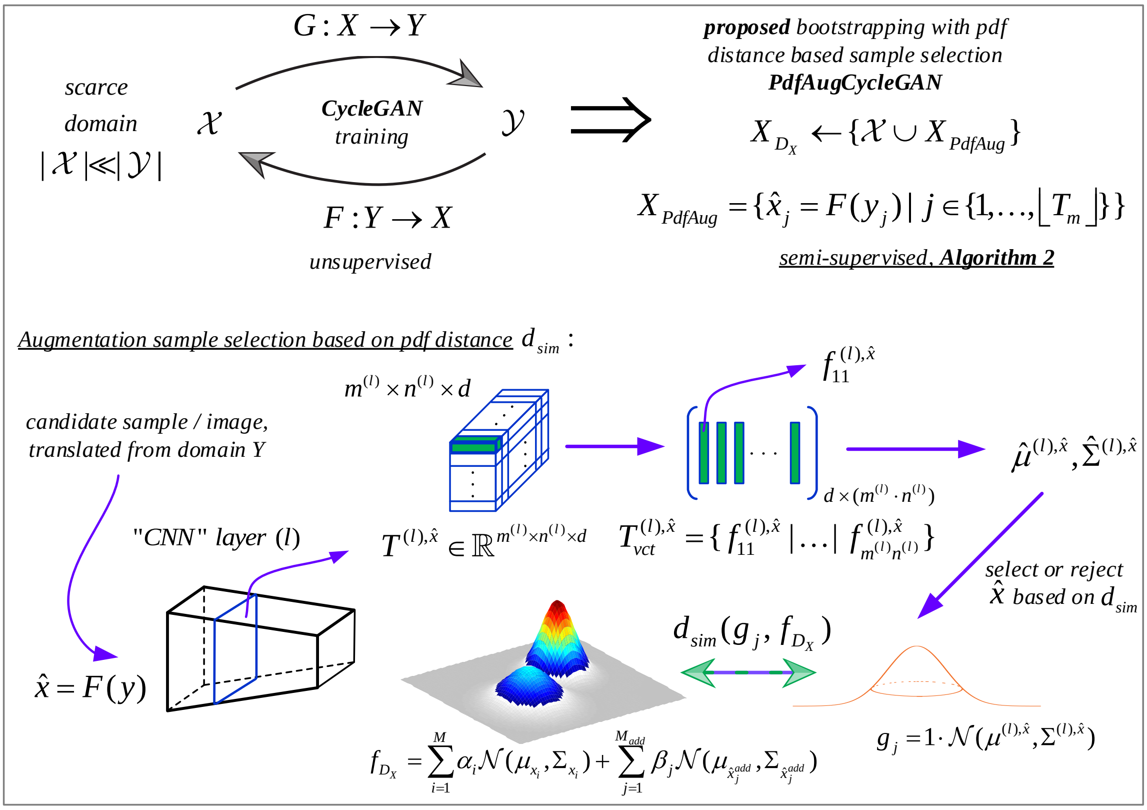

$G : X \to Y$
***proposed** bootstrapping with pdf distance based sample selection*
PdfAugCycleGAN
scarce domain
$|\mathcal{X}| \ll |\mathcal{Y}|$
$\mathcal{X}$
***CycleGAN** training*
$\mathcal{Y}$
$X_{D_X} \leftarrow \{\mathcal{X} \cup X_{PdfAug}\}$
$X_{PdfAug} = \{\hat{x}_j = F(y_j) \mid j \in \{1,\ldots,\lfloor T_m \rfloor\}\}$
$F : Y \to X$
unsupervised
*semi-supervised, **Algorithm 2***
Augmentation sample selection based on pdf distance d_{sim} :
$f_{11}^{(l),\hat{x}}$
$m^{(l)} \times n^{(l)} \times d$
candidate sample / image, translated from domain Y
$\hat{\mu}^{(l),\hat{x}}, \hat{\Sigma}^{(l),\hat{x}}$
$d \times (m^{(l)} \cdot n^{(l)})$
"CNN" layer (l)
$T^{(l),\hat{x}} \in \mathbb{R}^{m^{(l)} \times n^{(l)} \times d}$
$T_{vct}^{(l),\hat{x}} = \{ f_{11}^{(l),\hat{x}} | \ldots | f_{m^{(l)}n^{(l)}}^{(l),\hat{x}} \}$
select or reject $\hat{x}$ *based on* d_{sim}
$d_{sim}(g_j, f_{D_X})$
$\hat{x} = F(y)$
$g_j = 1 \cdot \mathcal{N}(\mu^{(l),\hat{x}}, \Sigma^{(l),\hat{x}})$
$f_{D_X} = \sum_{i=1}^{M} \alpha_i \mathcal{N}(\mu_{x_i}, \Sigma_{x_i}) + \sum_{j=1}^{M_{add}} \beta_j \mathcal{N}(\mu_{\hat{x}_j^{add}}, \Sigma_{\hat{x}_j^{add}})$

Figure 1. Illustration of the proposed *PdfAugCycleGAN* data augmentation process in Algorithm 2.

3.2. *PdfAugCycleGAN Training Procedure*

Algorithm 2 begins by performing the SSL strategy, during the first K_0 out of N iterations to avoid overfitting of the discriminator of the scarce domain D_X. The previously described augmentation strategy is then invoked, and it is performed periodically. Here we formalize the training procedure of the previously proposed *PdfAugCycleGAN* based on the following pseudo-code in Algorithm 2.

We note that the proposed mechanism for the selection of new augmented training samples first preselects only the percent P of translated examples that have pdfs in VGG19 feature space and are the closest ones to the GMM corresponding to the samples already in the training pool of the discriminator in the scarce domain. Thus, the reported performance improvement of *PdfAugCycleGAN* in comparison to the baseline *BTS-SSLCycleGAN* comes from the fact that translated examples that can be considered as outliers for the discriminator training are excluded from the training pool in the case of *PdfAugCycleGAN*. In contrast, in baseline *BTS-SSLCycleGAN*, all samples translated from the fully observable domain are always included in the scarce domain discriminator training.

Algorithm 2 PdfAugCycleGAN training

```
procedure PdfAugCycleGAN
    $N$—number of iterations; $K_0$—number of initial SSL iterations, $K_0 \ll N$; $K$—period of the proposed bootstrapping strategy (BTS) repetition; $M_{SSL} = |\mathcal{P}_{data}|$, number of paired samples in $\mathcal{P}_{data} = \{(x_i^p, y_i^p) | i = 1, \ldots, M_{SSL}\} = X_p \times Y_p$; $m_{SSL}$—SSL minibatch size, $m_{SSL} < M_{SSL}$; $m$—minibatch size after initial SSL phase; $P \in (0,1)$, percentage of unpaired samples translated from $Y$ to $X$ during the BTS minibatch that will be used to augment the original training pool corresponding to $D_X$; $P$ controls the number of selected translated examples with the lowest score $d_{sim}(g_j, f_{D_X})$ of all the translated examples $\hat{x}_j = G(y_j)$, which will be added to the training pool of $D_X$; $d_{sim}$ is a measure of similarity between GMMs; $\eta > 0$, learning rate; $X_u \subset X$, $Y_u \subset Y$, unpaired training subsets, while $X_p \subset X, Y_p \subset Y$ are paired training subsets, such that $|Y_p| \ll |Y_u|$, $|X_p| \ll |X_u|$, and $|X| \ll |Y|$;
    Randomly initialize the parameters of $D_X$, $D_Y$, and $G_{X \to Y}, G_{Y \to X}$: $\theta_{D_X}, \theta_{D_Y}, \theta_{G_{X \to Y}}, \theta_{G_{Y \to X}}$

    for $k = 1$ to $N$ do
        Sample minibatch of unpaired training data $\{x_1, \ldots, x_m\} \subset X_u$, $\{y_1, \ldots, y_m\} \subset Y_u$
        Sample minibatch of paired training data $\{(x_1^p, y_1^p), \ldots, (x_{m_{SSL}}^p, y_{m_{SSL}}^p)\} \subset X_p \times Y_p$
        $\mathcal{L}_1 = \widehat{\mathcal{L}}_{adv}(G_{Y \to X}, D_X) + \lambda_{cyc} \mathcal{L}_{cyc}(G_{X \to Y}, G_{Y \to X}) + \lambda_{SSL} \widehat{\mathcal{L}}_{SSL}(G_{X \to Y}, G_{Y \to X})$
        $\mathcal{L}_2 = \widehat{\mathcal{L}}_{adv}(G_{X \to Y}, D_Y) + \lambda_{cyc} \mathcal{L}_{cyc}(G_{X \to Y}, G_{Y \to X}) + \lambda_{SSL} \widehat{\mathcal{L}}_{SSL}(G_{X \to Y}, G_{Y \to X})$
        $\theta_{D_X}^{(k+1)} \longleftarrow \theta_{D_X}^{(k)} - \eta \nabla_{\theta_{D_X}} \widehat{\mathcal{L}}_{adv}(G_{Y \to X}, D_X)$
        $\theta_{D_Y}^{(k+1)} \longleftarrow \theta_{D_Y}^{(k)} - \eta \nabla_{\theta_{D_Y}} \widehat{\mathcal{L}}_{adv}(G_{X \to Y}, D_Y)$
        $\theta_{G_{Y \to X}}^{(k+1)} \longleftarrow \theta_{G_{Y \to X}}^{(k)} - \eta \nabla_{\theta_{G_{Y \to X}}} \mathcal{L}_1$
        $\theta_{G_{X \to Y}}^{(k+1)} \longleftarrow \theta_{G_{X \to Y}}^{(k)} - \eta \nabla_{\theta_{G_{X \to Y}}} \mathcal{L}_2$
        if $\big( (k > K_0) \wedge (( (k - K_0) \mod K) == 0) \big)$ then
            Perform the augmentation of the training pool of discriminator $D_X$:
            Sort all translated examples $\hat{x}_j = G_{Y \to X}(y_j), j \in \{1, \ldots, m\}$, by the increasing values of $d_{sim}(g_j, f_{D_X})$, so that $\{\hat{x}_j = G_{Y \to X}(y_j)\}$ is sorted in that manner, then select the first $\lfloor Tm \rfloor$ that correspond to the percent $P$ of the translated samples from $Y_u$, i.e., $T_m = P|Y_u|$, and extend the training pool of $D_X$:
            $X_{D_X} \leftarrow \{X \cup X_{PdfAug}\}$, $X_{PdfAug} = \{\hat{x}_j = F(y_j) | j \in \{1, \ldots, \lfloor Tm \rfloor\}\}$
        end if
    end for
end procedure
```

3.3. Measure of Similarity Between GMM Mixtures Used in Data Augmentation Process

The KL divergence, defined as $KL(p||q) = \int_{\mathbb{R}^d} p(x) \ln \frac{p(x)}{q(x)} dx$, is the most natural measure between two probability distributions p and q. For the proposed pdf distance mentioned in the previous section, d_{sim}, we use the GMM similarity measure based on KL divergence between GMMs. Since it does not exist in closed form, an approximation based on the closed form expression for KL divergence between corresponding multivariate Gaussian components can be given by Equation (15).

Let us denote two GMMs as $f = \sum_{i=1}^{n} \alpha_i f_i$ and $g = \sum_{j=1}^{m} \beta_j g_j$, with $f_i = \mathcal{N}(\mu_{f_i}, \Sigma_{f_i})$ and $g_j = \mathcal{N}(\mu_{g_j}, \Sigma_{g_j})$ representing Gaussian components of the corresponding mixtures, with weights $\alpha_i \geq 0$, $\beta_j \geq 0$ $\sum_{i=1}^{n} \alpha_i = 1$, $\sum_{j=1}^{m} \beta_j = 1$. Terms μ_{f_i}, μ_{g_j} are means, while $\Sigma_{f_i}, \Sigma_{g_j}$ are covariance matrices of f_i and g_j. Then, the KL divergence between two Gaussian components $KL(f_i||g_j)$ exists in the closed form given as follows:

$$KL(f_i||g_j) = \ln \frac{|\Sigma_{f_i}|}{|\Sigma_{g_j}|} + Tr\left[\Sigma_{g_j}^{-1}\Sigma_{f_i}\right] + (\mu_{f_i} - \mu_{g_j})^T \Sigma_{g_j}^{-1} (\mu_{f_i} - \mu_{g_j}) - d. \tag{14}$$

Thus, the roughest approximation for KL divergence between GMMs, based on the convexity of the KL divergence, is given as follows:

$$KL(f||g) \leq KL_{WA}(f||g) = \sum_{i,j} \alpha_i \beta_j KL(f_i||g_j), \tag{15}$$

where $KL(f_i||g_j), i = 1, \ldots, n, j = 1, \ldots, m$ are given by (14).

For this paper, we use the approximation of the KL divergence between GMMs f and g based on averaging. Namely, GMMs f and g are replaced with multivariate Gaussians $\hat{f} = \mathcal{N}(\mu_{\hat{f}}, \Sigma_{\hat{f}})$ nad $\hat{g} = \mathcal{N}(\mu_{\hat{g}}, \Sigma_{\hat{g}})$ with

$$\begin{aligned} \mu_{\hat{f}} &= \sum_i \alpha_i \mu_{f_i} \\ \Sigma_{\hat{f}} &= \sum_i \alpha_i \left(\Sigma_{f_i} + (\mu_{f_i} - \mu_{\hat{f}})(\mu_{f_i} - \mu_{\hat{f}})^T \right). \end{aligned} \tag{16}$$

and similarly for g and $\hat{g}$. Those estimates for $\hat{f}$ and $\hat{g}$ obtain minimum $KL(\tilde{f}||\tilde{g})$, with $\tilde{f}, \tilde{g}$ in the class of multivariate Gaussians with a predefined dimension. Thus, the KL divergence between f and g is approximated by $KL(f||g) \approx KL(\hat{f}, \hat{g})$, where $KL(\hat{f}, \hat{g})$ is evaluated using (14).

4. Experimental Results

In this section, we present the experiments obtained on several real-world datasets, in the problem of image translation: (1) Semantic label ↔ photo task on *CityScapes* dataset [11,39]. This dataset consists of 2975 training images of the size 128 × 128, as well as an evaluation set for testing; (2) Architectural labels ↔ photo task [11,40] on the CMP *Facade dataset*, containing 400 training images; (3) Map ↔ aerial photo task on *Google Maps* dataset [11], containing 1096 training images of the size 256 × 256.

The experimental setup for all datasets was designed to simulate an imbalanced, i.e., scarce domain scenario, where the left or the target domain X in the image translation task is considered scarce (with significantly less training data in comparison to the source domain Y). This is achieved using a setup in which only a certain percentage of the original left domain is used for I2I model training.

Each of the described experiments compares the proposed *PdfAugCycleGAN* image translation method against the following baseline methods: (1) original unsupervised CycleGAN proposed in [15], (2) fully supervised *pix2pix* method proposed in [11], (3) semi-supervised *BTS-SSLCycleGAN* method proposed in [20].

In order to compare the performance over the aforementioned tasks and datasets, we used Peak Signal-to-Noise Ratio (PSNR), as well as the more advanced Structural Similarity Index Measure (SSIM), which is a more advanced perception-based model that considers image degradation as perceived change in image structural information. The PSNR is evaluated as $PSNR = 20log\left(MAX_I / \sqrt{MSE}\right)$, where MAX_I is the maximum possible pixel value of the ground truth images, while MSE is the squared Euclidean norm between the generated and ground truth images. The SSIM measure between generated images and ground truth images is calculated on various windows x, y of an image, as follows:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{17}$$

where μ_x and μ_y are the average values of x and y, while σ_x^2 and σ_y^2 are variances and c_1 and c_2 are constants, set as reported in [41].

4.1. Experimental Setup

Through all experiments (all datasets, all baselines, as well as the proposed method), we varied the percentage of the data that was considered as available in the scarce domain in the following steps: 25%, 50%, and 100%.

For the baseline *BTS-SSLCycleGAN* and the proposed *PdfAugCycleGAN*, in all experiments, after the initial K_0 iterations, during the next $k + K$ training iterations, where $K = 50$, the examples from the fully observable (original) domain were transformed by $G_{Y\to X}$ and added to the training pool of the discriminator D_X to perform the proposed data augmentation strategy (BTS). This periodical bootstrapping of the scarce domain during model training was done such that 20% of randomly chosen examples (by uniform distribution) was generated for *BTS-SSLCycleGAN* in each BTS iteration.

For the proposed *PdfAugCycleGAN*, in each BTS iteration (on every K iteration of model training after K_0) the fixed percent of 50% ($p = 0.5$) of examples translated by $G_{Y\to X}$ with the lowest $d_{sim}(g_j, f_{D_X})$ score was selected for augmentation of the D_X training pool, where g_j, f_{D_X}, and d_{sim} are defined in Section 3.3 (these were the percents for which we obtained the best results). The described bootstrapping procedure was repeated (with the same rate as in the baseline *BTS-SSLCycleGAN* method), and new translated samples were periodically added into the training pool of the scarce domain X during the proposed CycleGAN model training strategy with sample selection based on pdf distance.

Since all the considered datasets originally contained paired images, we also used a fixed 20% of paired training examples from the scarce domain for the initial semi-supervised learning stage of both *BTS-SSLCycleGAN* and *PdfAugCycleGAN*. The rest of the available data in the scarce domain were prepared in such a way that their corresponding pairs from the original dataset were discarded from the target domain in each of the experiments. In the asymmetric I2I translation task that was of the most interest for the proposed data augmentation method (scenario in which the total sample size of the scarce domain corresponds to only 25% of original paired data), the result of the described setup was that only a relatively small amount of paired data was available for the SSL stage (5% of the original dataset), while the remaining 20% of the available (unpaired) data in the scarce domain were reserved for the investigation of the proposed unsupervised model training strategy. Through sample selection involving pdf distance computation, Algorithm 2, the scarce domain was then periodically extended during the *PdfAugCycleGAN* model training procedure.

Considering the actual CycleGAN generator network architecture, we used the one originally proposed in [15] and also utilized in [17], as well as in [20]. It contains two stride-2 convolutions, several residual blocks, and two fractionally strided convolutions (stride $\frac{1}{2}$). It also utilizes six blocks for 128 × 128 and nine blocks for 256 × 256 and higher-resolution type images and also instance normalization, as in [15,42,43]. Considering the discriminator network, we used 70 × 70 PatchGAN.

Considering the VGG19 network used in the evaluation of image feature maps and the construction of corresponding pdfs g_j, f_{D_X}, on which the computation of the selection score

$d_{sim}(g_j, f_{D_X})$ is based, we use the standard pretrained VGG19, as described in Section 3.1 (see also [26]).

Considering the training procedure, instead of the loss function in Equation (3), we used a more stable L_2 adversarial loss (as reported in [44]). We also use the history of 50 generated images in order to calculate the average score. For all experiments, we used $\lambda_{cyc} = \lambda_{SSL} = 10$, $m = m_{SSL} = 50$ with a learning rate $\eta = 0.0002$, which was kept constant during the first 100 epochs, linearly decaying to zero during the next 100 epochs. Network parameters were initialized by using random samples drawn from the normal distribution $\mathcal{N}(0, 0.02)$.

4.2. Result Analysis and Discussion

In Table 1, the experimental results of the proposed *PdfAugCycleGAN* in comparison to the baseline pix2pix, CycleGAN, and *BTS-SSLCycleGAN* are presented in terms of PSNR and SSIM measures on several databases.

It can be seen that, in the majority of experiments involving unpaired (unsupervised) or semi-supervised I2I translation tasks, the proposed *PdfAugCycleGAN* obtained improvements in comparison to the baseline *BTS-SSLCycleGAN*, as well as the classical CycleGAN method, in both PSNR as well as SSIM measures. More specifically, in experiments obtained on *CityScapes* and *Facade datasets*, the proposed *PdfAugCycleGAN* obtained better results in comparison to all baseline methods, while in experiments on the *Google Maps* dataset, the *BTS-SSLCycleGAN* performed slightly better, i.e., for that particular dataset, the proposed selective bootstrapping methodology failed to improve the I2I translation results. However, the proposed *PdfAugCycleGAN* did achieve better PSNR performance on the particular dataset in the I2I translation scenario with the smallest sample size of the scarce domain X.

In general, the reported PSNR advantage of *PdfAugCycleGAN* over *BTS-SSLCycleGAN* is always present under the challenging scenario of a small sample size of X (when only 25% of dataset samples was considered as available for the experiment), and even higher in the case of other I2I tasks and datasets.

Table 1. Experimental comparison between proposed *PdfAugCycleGAN* and the baseline *CycleGAN*, *BTS-SSLCycleGAN*, and fully supervised *pix2pix*, under different scenarios: varying sample size $\mathcal{S}_X$ of scarce domain $\mathcal{X}$, as well as when applied to different tasks/datasets—*CityScapes*, *Facade dataset*, *Google Maps*. Bold values indicate better performance among the last two methods.

	$\mathcal{S}_X$	pix2pix		CycleGAN		BTS-SSLCycleGAN		PdfAugCycleGAN	
	[%]	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
CityScapes	25	19.89	0.60	17.20	0.56	18.77	0.61	**19.10**	**0.64**
	50	20.45	0.64	17.00	0.55	19.04	0.64	**19.22**	**0.67**
	100	19.51	0.59	17.12	0.54	20.47	0.65	**21.23**	**0.68**
Facade dataset	25	13.78	0.35	10.93	0.25	11.83	0.32	**12.14**	**0.34**
	50	14.24	0.40	11.00	0.25	13.75	0.40	**13.92**	**0.42**
	100	14.25	0.42	10.98	0.27	13.21	0.41	**13.57**	**0.43**
Google Maps	25	30.35	0.67	30.47	0.71	**31.20**	**0.77**	30.90	0.76
	50	30.55	0.68	29.78	0.72	**30.88**	**0.79**	30.21	0.76
	100	30.01	0.69	30.24	0.73	**31.23**	**0.81**	30.89	0.79

Besides the unfavorable I2I translation scenario with asymmetric sample size, where *PdfAugCycleGAN* outperformed the competing BTS-SSL on both *CityScapes* and *Facade datasets* (thanks to the proposed selective approach to data augmentation), in the case of the same sample size ($\mathcal{S}_X = 100\%$ in Table 1), *PdfAugCycleGAN* also achieved better performance in comparison to the fully supervised *pix2pix* method, which further justifies the proposed selective BTS strategy.

Overall, the results in Table 1 confirm that the proposed more subtle handling of the translated examples, concerning whether those should be added into the pool of the discriminator D_X of the scarce domain, can improve the performance of the augmented CycleGAN system.

In Figure 2, visual examples are given for the proposed *PdfAugCycleGAN* vs. baseline *CycleGAN*, semi-supervised *BTS-SSLCycleGAN*, and fully supervised *pix2pix* algorithm comparisons for 50% of the scarce domain data used. Real A (fully observable domain) and Real B (scarce, target domain) correspond to image pair examples. Examples are shown for *Google Maps*, *CityScapes*, and *Facade datasets*. It can be seen that in these examples, the proposed *PdfAugCycleGAN* obtains visually more accurate results than the baseline methods.

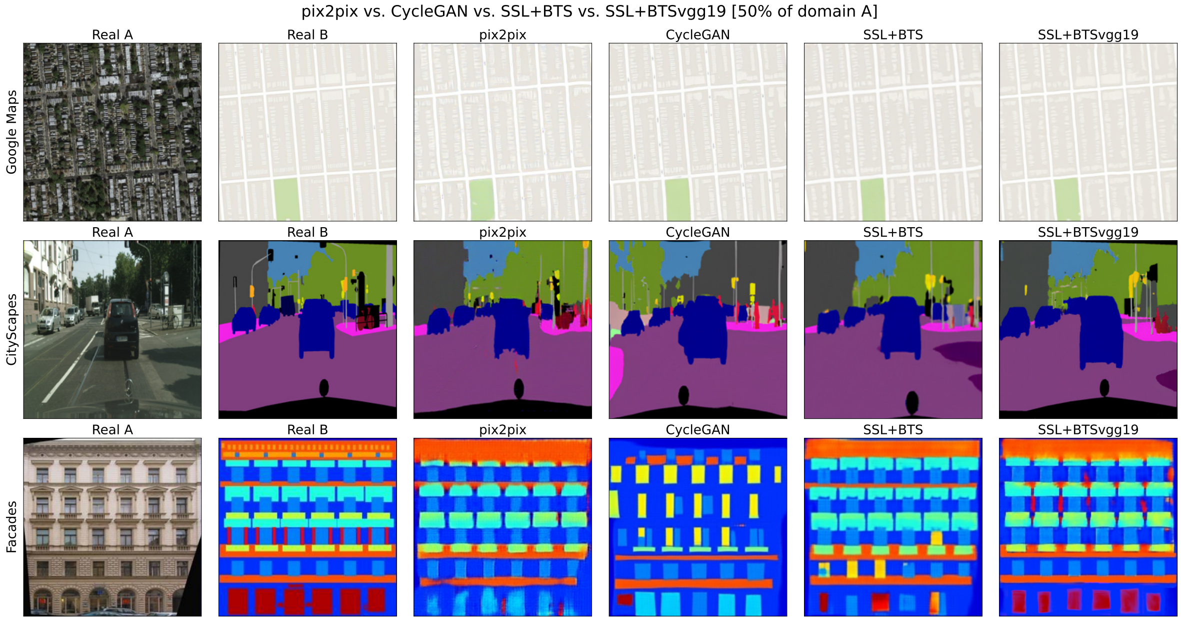


Figure 2. Visual examples are given for the proposed *PdfAugCycleGAN* vs. baseline textitCycleGAN, *BTS-SSLCycleGAN*, and fully supervised *pix2pix* algorithm comparisons for 50% of the scarce domain data used. Real A (fully observable domain) and Real B (scarce, target domain) correspond to image pair examples. Results of *PdfAugCycleGAN* are denoted by "SSL + BTSvgg19", and similarly for *BTS-SSLCycleGAN* by "SSL + BTS" image labels.

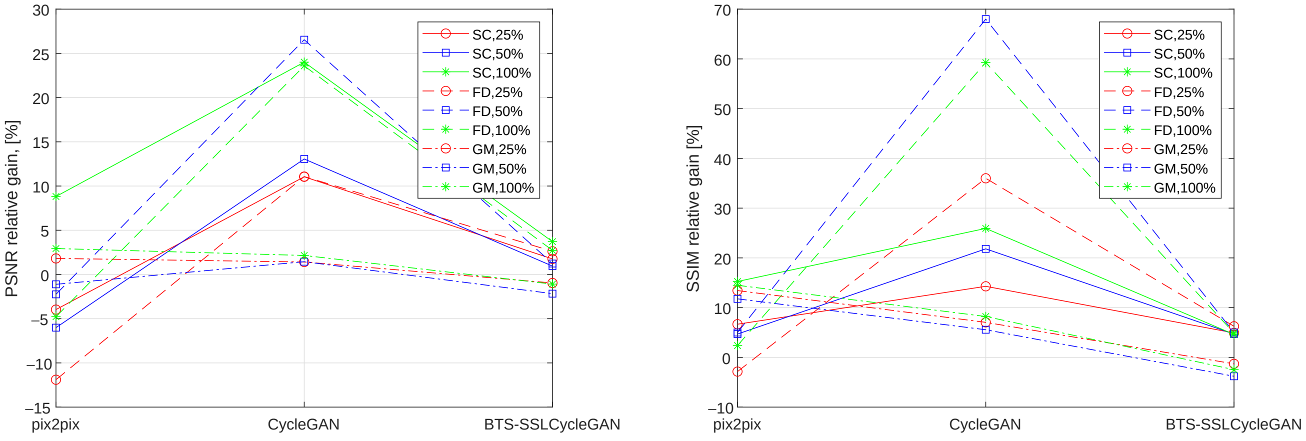
Based on additional analysis of reported values of PSNR [dB] and SSIM in Table 1, which are shown in Figure 3 we can also observe the following.

The relative gain γ of the proposed *PdfAugCycleGAN* is consistently positive in comparison to competing baseline methods (*CycleGAN* and *BTS-SSLCycleGAN*) when measured over *CityScapes* and *Facade datasets* under all training data conditions (25%, 50%, and 100% of available training data). It is computed as the normalized difference of the performance metric values:

$$\gamma = \frac{\zeta(PdfAugCycleGAN) - \zeta(baseline)}{\zeta(baseline)} \cdot 100\% \,, \tag{18}$$

where $\zeta(\cdot)$ denotes either PSNR or SSIM metric, i.e., performance values corresponding to the proposed *PdfAugCycleGAN* and selected baseline method of interest.

From the presented results, it is also possible to see that this gain is not significantly lower in the case of the 25% scenario in comparison to the 50% and 100% experiments in the case of comparisons with the *CycleGAN* and *BTS-SSLCycleGAN* methods. However, in the case of certain experiments where the evaluation of the proposed method is performed over the *Google Maps* dataset, the obtained relative gain against *BTS-SSLCycleGAN* was negative, as shown in Figure 3. The values indicate that the proposed method, for *CityScapes* and *Facade datasets*, provides consistent improvement with respect to *BTS-SSLCycleGAN*, including a relative increase of PSNR in the range of 1.2–3.7%, and a relative increase of SSIM in the range of 4.6–6.2%. Losses are observed for the *Google Maps* dataset, less than 2.2% for PSNR and less than 3.8% for SSIM. Slightly lower performance metrics on the *Google Maps* dataset can be explained by the fact that the proposed augmented model training is utilizing sample selection that is dependent on VGG19 feature space. This means that pdf estimation is relying on a feature extraction CNN that is pre-trained on natural image scenes, which are completely different from the satellite image scenes present in the *Google Maps* dataset. Thus, the proposed pdf distance computation is to some extent dependent on the type of images (scenes) on which the pre-trained feature extraction network is trained. Therefore, the proposed adaptive augmentation of the scarce domain in *PdfAugCycleGAN* is affected by the characteristics of the selected pre-trained feature space (character of the side information). This leads to better results over *CityScapes* and *Facade datasets* in comparison to *Google Maps*, which has a different type of scenes (significantly different from the ones on which VGG19 was trained).



(**a**) (**b**)

Figure 3. Relative gain γ of *PdfAugCycleGAN* against competing baseline methods in terms of: (**a**) PSNR, and (**b**) SSIM measures, which are reported in Table 1. Specific markers (colors) denote training scenario: 25%, 50%, or 100% of available data; while different line types define dataset type: *CityScapes*, *Facade dataset*, or *GoogleMaps*.

At the end, we note that other limitations could also come from the choice and type of the training datasets and methods, e.g., the authors in [21,22] use dataset types which are problem-specific and significantly different from the ones used in the presented experiments (presence of cross-modal semantic information or clearly distinguishable classes), as already discussed in Section 1.

5. Conclusions

In this paper, we have proposed a novel approach for the CycleGAN training strategy in the case of unfavorable domain translation scenarios in which the original data domain has a substantially smaller number of samples in comparison to the target data domain. The method is based on periodical boosting of scarce domain statistics through data augmentation based on selective sampling of novel data samples generated via domain translation from the fully observable domain. The described data augmentation and learning are performed in a fully unsupervised manner, after a short initial semi-supervised stage that prevents the overfitting of the discriminator network in the scarce data domain. This process is periodically repeated throughout the model training process and significantly improves the overall model performance in comparison to the unsupervised learning baseline. In comparison to the similar previously proposed semi-supervised learning strategy, which also relies on data augmentation through transfer of samples from the fully observable domain, the proposed method achieves selective bootstrapping and thus better performance in cases when the sample size in the scarce domain is several times smaller than the number of samples in the fully observable domain.

In addition to the training strategy, this paper also proposes a novel sample selection criterion based on pdf distance between distributions of feature vectors corresponding to learned image representations obtained by a pretrained CNN. Thus, the proposed similarity measure between the specific sample and the pool of existing samples in the system could also apply to other types of problems involving adaptive sample selection or distance computation in the learned feature space characteristic for some specific signal modality.

The presented results also confirm that selective or more subtle strategies for data augmentation could be a key for more efficient model learning in the case of imbalanced domain translation tasks.

Author Contributions: Conceptualization, L.K., B.P., B.B., M.N., and S.B.; methodology, L.K., B.P., B.B., M.N., and S.B.; software, B.P.; validation, B.P., M.N., and L.K.; formal analysis, B.P., L.K., B.B., and M.N.; investigation, L.K., B.P., and B.B.; resources, B.P.; data curation, B.P.; writing—original draft preparation, L.K., B.P., B.B., and S.B.; writing—review and editing M.N., B.B., and L.K.; visualization, B.B., and B.P.; supervision, L.K., and B.P.; project administration, L.K., and S.B.; funding acquisition, L.K., and B.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Science Fund of the Republic of Serbia, project "Multimodal multilingual human-machine speech communication—AI SPEAK", grant no. 7449; by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (contract no. 451-03-137/2025-03/200156); and by the Faculty of Technical Sciences, University of Novi Sad, through the project “Scientific and artistic research work of researchers in teaching and associate positions at the Faculty of Technical Sciences, University of Novi Sad 2025” (no. 01-50/295).

Data Availability Statement: Publicly available datasets were analyzed in this study. *Google Maps* data can be found here: [http://efrosgans.eecs.berkeley.edu/cyclegan/datasets/maps.zip (accessed on 14 December 2024)]. CMP *Facade dataset* data can be found here: [http://efrosgans.eecs.berkeley.edu/cyclegan/datasets/facades.zip (accessed on 14 December 2024)]. *CityScapes* data are available on request from: [https://cityscapes-dataset.com (accessed on 14 December 2024)].

Acknowledgments: B.P., M.N., and B.B. would like to acknowledge the support of the Faculty of Technical Sciences through the project: “Development and application of modern research tools and methods in power, electronic and telecommunication engineering”.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

BTS	Bootstrapping strategy
BTS-SSL	Bootstrapped semi-supervised learning
BTS-SSLCycleGAN	BTS-SSL CycleGAN
GAN	Generative adversarial network
cGAN	Conditional GAN
CNN	Convolutional neural network

CycleGAN	Adaptive cycle-consistent GAN
EM	Expectation maximization
GMM	Gaussian mixture model
I2I	Image-to-image domain translation
KL	Kullback–Leibler
ML	Maximum likelihood method
MSE	Mean squared error
pdf	Probability density function
PdfDistCycleGAN	pdf distance-based augmented CycleGAN
PSNR	Peak signal-to-noise ratio
SSL	Semi-supervised learning
SSIM	Structural similarity index measure

Appendix A

In the following, we provide a list of the main symbols and their descriptions:

$x \in X$	sample from the target or "left", scarce domain X
$y \in Y$	sample from the original or "right", fully observable domain Y
G, F	generator networks operating in domains X and Y, respectively
$D, D_X(\cdot), D_Y(\cdot)$	discriminator networks (for domains X and Y)
$p(\cdot)$	data distribution
$\mathbb{E}_{x \sim p(x)}$	mathematical expectation over $p(x)$
$G_{X \to Y}, G_{Y \to X}$	direct and the inverse I2I translations (nonlinear generator mappings)
$\mathcal{L}_{adv}(\cdot,\cdot)$	adversarial loss function
$\mathcal{L}_{cycle}(\cdot,\cdot)$	cycle consistency loss function
$\mathcal{L}_{SSL}(\cdot,\cdot)$	semi-supervised loss function
$\mathcal{L}(G, F, D_X, D_Y)$	CycleGAN learning objective (with or without SSL loss)
$\lambda_{cycle},\ \lambda_{SSL}$	weights of regularization terms
$\mathcal{P}$	set of indices of paired samples from the original and target domains
$\\| \cdot \\|_{l_p}$	l_p vector norms
$\theta^{(k)}$	model parameters in training iteration k
d	feature space dimensionality in $\mathbb{R}^d$
$T^{(l),\hat{x}}$	feature map tensor from layer l of CNN processing translated image $\hat{x}$
$m^{(l)} \times n^{(l)}$	spatial dimensions of feature map in l-th layer of CNN
$T_{vct}^{(l),\hat{x}}$	vectorized feature map tensor (matrix with $m^{(l)}n^{(l)}$ columns of size d)
$\mathcal{N}(\mu, \Sigma)$	multivariate normal distribution with mean μ and covariance Σ
$KL(\cdot\\|\\|\cdot)$	Kullback–Leibler divergence between normal distributions
$\mathcal{N}(\mu_{\hat{x}_j}, \Sigma_{\hat{x}_j})$	pdf of CNN features from the translated sample $\hat{x}_j$
$\hat{\mu}^{(l),\hat{x}}$	ML estimate of pdf mean based on $\mathbb{R}^d$ CNN feature vectors from $T_{vct}^{(l),\hat{x}}$
$\hat{\Sigma}^{(l),\hat{x}}$	sample covariance matrix based on $\mathbb{R}^d$ CNN feature vectors from $T_{vct}^{(l),\hat{x}}$
f_{D_X}	pdf of CNN feature vectors of samples in the training pool of D_X
M	number of original image samples in the training pool of D_X
M_{add}	number of added image samples to scarce domain X through BTS
f_i, g_j	components in GMMs f and g; $g_j = \mathcal{N}(\mu_{g_j}, \Sigma_{g_j})$, $f_i = \mathcal{N}(\mu_{f_i}, \Sigma_{f_i})$
$d_{sim}(\cdot,\cdot)$	distance function between pdfs
$SSIM(\cdot,\cdot)$	structural similarity index measure between images

References

1. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2337–2346.
2. Zhu, P.; Abdal, R.; Qin, Y.; Wonka, P. SEAN: Image synthesis with semantic region-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5104–5113.
3. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. MaskGAN: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5549–5558.
4. Tang, H.; Xu, D.; Yan, Y.; Torr, P.H.; Sebe, N. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7870–7879.
5. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. DualGAN: Unsupervised dual learning for image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2849–2857.
6. Mejjati, Y.A.; Richardt, C.; Tompkin, J.; Cosker, D.; Kim, K.I. Unsupervised attention-guided image-to-image translation. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 3697–3707.
7. Tomei, M.; Cornia, M.; Baraldi, L.; Cucchiara, R. Art2Real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5849–5859.
8. Liu, M.Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; Kautz, J. Few-shot unsupervised image-to-image translation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10551–10560.
9. Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 701–710.
10. Zhang, Y.; Liu, S.; Dong, C.; Zhang, X.; Yuan, Y. Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE Trans. Image Process.* **2019**, *29*, 1101–1112. [CrossRef]
11. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
12. Wang, C.; Zheng, H.; Yu, Z.; Zheng, Z.; Gu, Z.; Zheng, B. Discriminative region proposal adversarial networks for high-quality image-to-image translation. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 770–785.
13. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional GANs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
14. AlBahar, B.; Huang, J.B. Guided image-to-image translation with bi-directional feature transformation. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9016–9025.
15. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
16. Bao, F.; Neumann, M.; Vu, N.T. CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition. In Proceedings of the Conference of the International Speech Communication Association (Interspeech), Graz, Austria, 15–19 September 2019; pp. 2828–2832.
17. Meng, Z.; Li, J.; Gong, Y.; Juang, B.H. Cycle-consistent speech enhancement. In Proceedings of the Conference of the International Speech Communication Association (Interspeech), Hyderabad, India, 2–6 September 2018; pp. 1165–1169.
18. Hosseini-Asl, E.; Zhou, Y.; Xiong, C.; Socher, R. Robust domain adaptation by augmented cyclic adversarial learning. In Proceedings of the 31st international Conference on Neural Information Processing Systems-Interpretability and Robustness for Audio, Speech and Language Workshop, Montreal, QC, Canada, 3–8 December 2018; pp. 1–11.
19. Qi, C.; Chen, J.; Xu, G.; Xu, Z.; Lukasiewicz, T.; Liu, Y. SAG-GAN: Semi-supervised attention-guided GANs for data augmentation on medical images. *arXiv* **2020**, arXiv:2011.07534.
20. Krstanovic, L.; Popovic, B.; Janev, M.; Brkljac, B. Bootstrapped SSL CycleGAN for Asymmetric Domain Transfer. *Appl. Sci.* **2022**, *12*, 3411. [CrossRef]
21. Deng, S.; Uchida, K.; Yin, Z. Cross-modal and semantics-augmented asymmetric CycleGAN for data-imbalanced anime style face translation. In Proceedings of the 3rd International Conference on Video, Signal and Image Processing, Wuhan, China, 19–21 November 2021; pp. 43–51.

22. Patashnik, O.; Danon, D.; Zhang, H.; Cohen-Or, D. BalaGAN: Cross-modal image translation between imbalanced domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2659–2667.
23. Gladh, M.; Sahlin, D. Image Synthesis Using CycleGAN to Augment Imbalanced Data for Multi-Class Weather Classification. Master's Thesis, Department of Science and Technology, Linköping University, Faculty of Science & Engineering, Linköping, Sweden, 2021.
24. Yang, X.; Wang, L. SDP-CycleGAN: CycleGAN based on siamese data pairs for unraveling data imbalance problems in anomaly detection. In Proceedings of the 9th International Conference on Advanced Cloud and Big Data (CBD), IEEE, Xi'an, China, 26–27 March 2022; pp. 151–157.
25. Kim, D.; Byun, J. Data augmentation using CycleGAN for overcoming the imbalance problem in petrophysical facies classification. In Proceedings of the SEG International Exposition and Annual Meeting, SEG, Online, 11–16 October 2020; p D031S041R004.
26. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
27. Hershey, J.R.; Olsen, P.A. Approximating the Kullback Leibler divergence between Gaussian mixture models. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. 317–320.
28. Hao, H.; Wang, Q.; Li, P.; Zhang, L. Evaluation of ground distances and features in EMD-based GMM matching for texture classification. *Pattern Recognit.* **2016**, *57*, 152–163. [CrossRef]
29. Goldberger, J.; Gordon, S.; Greenspan, H. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In Proceedings of the International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 3, pp. 487–493.
30. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis. (IJCV)* **2015**, *115*, 211–252. [CrossRef]
31. ImageNet. Available online: http://www.image-net.org (accessed on 2 December 2024).
32. Gatys, L.; Ecker, A.; Bethge, M. A Neural Algorithm of Artistic Style. *J. Vis. Sept.* **2016**, *16*, 326. [CrossRef]
33. An, J.; Xiong, H.; Huan, J.; Luo, J. Ultrafast Photorealistic Style Transfer via Neural Architecture Search. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, AAAI-20 Technical Tracks 7.
34. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M.-H. Universal Style Transfer via Feature Transforms. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 385–395.
35. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
36. Mirza, M. Conditional Generative Adversarial Nets. *arXiv* **2014**, arXiv:1411.1784v1.
37. Anderson, T.W.; Olkin, I. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear Algebra Its Appl.* **1985**, *70*, 147–171. [CrossRef]
38. Webb, A.R. *Statistical Pattern Recognition*; John Wiley & Sons: Hoboken, NJ, USA, 2011.
39. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
40. Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; Lee, H. Generative adversarial text to image synthesis. In Proceedings of the International Conference on Machine Learning, PMLR, New York, NY, USA, 20–22 June 2016; pp. 1060–1069.
41. Wang,Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600. [CrossRef] [PubMed]
42. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2017**, arXiv:1607.08022v3.
43. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 694–711.
44. Mao, X.; Li, Q.; Xie, H.; Lau, R.Y.K.; Wang, Z.; Smolley, S.P. Least Squares Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2813–2821.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.