COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE SESSION
INVITED PAPER

STYLE ADAPTATION BASED ON IMAGE PROCESSING METHODS USING CYCLEGAN

Branislav Popović*

Faculty of Technical Sciences,
University of Novi Sad,
Novi Sad, Serbia

Abstract:

Cycle-Consistent Generative Adversarial Networks (CycleGANs) are able to provide a highly under-constrained mapping between input and output data samples, i.e., source and target data domain, in cases when the aligned dataset is unavailable, in an unsupervised training fashion, using cycle-consistency loss mechanisms. On the other hand, most image-to-image and speech-to-speech translation tasks use the aligned, i.e., paired input-output training datasets. A large amount of data is necessary to train such architectures, while one of the domains could be scarce.

Several possible improvements to the original CycleGAN architecture are analysed in this paper for the cases when only a small percentage of training samples are aligned among source and target data domains. A semi-supervised approach is proposed to achieve better translation accuracy and prevent overfitting of the scarce data domain discriminator during initial training iterations. The training database is augmented by adding samples generated by inverse CycleGAN mappings after several training epochs (when the network is sufficiently trained) into the training pool of the discriminator of scarce, i.e., reduced data domain. An additional optimization constraint is also proposed, aligning probability distributions of feature maps belonging to the same-depth neural network layers of direct GAN encoder and inverse GAN decoder, to reinforce resemblance among object representations in various data domains.

Significantly better performances are obtained using proposed improvements in both image-to-image and speech-to-speech translation tasks, by observing standard qualitative and quantitative measures, in comparison to the baseline CycleGAN training approach.

Keywords:

Style Adaptation, Generative Adversarial Networks, Cycle-Consistency, Semi-Supervised Learning, Bootstrapping.

INTRODUCTION

Style transfer is a machine-learning technique presuming the translation of a particular referent style (e.g., painting technique, seasonal landscape features, colour schemes, etc. in case of images [1], or speaker characteristics, speaker emotion, pronunciation style, prosody, etc. in case of speech [2]) from one data sample to another (source to target domain), preserving at the same time the core attributes (content, structural features, semantics) of the original sample. Supervised learning methods use pairs of training data samples to be able to learn a one-direction

Correspondence:
Branislav Popović

e-mail:
bpopovic@uns.ac.rs

DOI: 10.15308/Sinteza-2023-9-16

sample-to-sample mapping among samples containing the same structural information. Many of those methods have originated from the conditional GAN (cGAN) training method [3], able to incorporate supplementary information (e.g., class labels), learning the difference between any two particular samples presenting the same entity in various domains.

Pix2Pix training architecture represents a fully-supervised cGAN-based training strategy comprising a U-Net-based generator [4] and a convolutional PatchGAN discriminator [5] able to cope with an extensive range of image-to-image translation tasks. Pix2Pix operates on real data in conjunction with labels in order to acquire mapping from the source to the target domain along with the reconstruction loss function, using pairs of one-to-one corresponding image representations from both domains. Albeit highly efficient, Pix2Pix cannot easily capture complex scene structural correlations using a single mapping, i.e., a single translation network (one generator and one discriminator). Furthermore, it is often difficult to aggregate a sufficient quantity of paired domain-to-domain training data to be able to train the network assuming appropriate precision and robustness. The latter is also a drawback of other supervised learning methods, such as DRPAN [6], ASAPNet [7] and SPADE [8].

Unsupervised learning methods, such as CycleGAN [9], CoGAN [10], DiscoGAN [11] or UNIT [12], learn corresponding mappings using sets of unaligned training samples from source and target domains (commonly highly under-constrained one-to-many-approach). CycleGAN employs two GANs working in opposite directions (each one has one generator and one discriminator able to provide mappings from source to target domain and back), and a cycle-consistency loss combined with adversarial loss, imposing bijection (mappings become reverses of each other by enforcing structural similarity between the original, i.e., source, and translated data samples after both forward and backward procedures have been completed). However, CycleGANs are unable to perform complex geometrical transformations and they are prone to diminish gradient issues and other types of instabilities, e.g., there are observable disparities in performances of the supervised (trained using pairs of one-to-one corresponding data samples) and the unsupervised version (trained using unaligned data).

Triggered as a result of a deficient number of structurally correlated samples between the source and the target domain (in the case when one of the domains is scarce), a bootstrapped semi-supervised BTS SSL CycleGAN algorithm is proposed [1]. Semi-supervised learning (SSL) strategy exploits the advantage of having a certain percentage of the aligned data samples in the training database to increase the accuracy and improve the overall performance of the CycleGAN algorithm, and at the same time prevents overfitting of both generator and discriminator related to the scarce domain using the rest of data in an unsupervised manner. The second step presumes periodical insertion (i.e., bootstrapping) of samples artificially produced by the generator related to the fully observable domain to the original training pool of the discriminator representing the scarce domain, but only after several training iterations have already been completed (the generator is adequately trained).

Feature Map Regularised FMR CycleGAN approach [13] adds an additional cycle-consistency loss to the objective function. The loss is calculated between probability density functions (PDFs) representing feature-map statistics of the same-depth neural network layers of the direct GAN encoder and the inverse GAN decoder to increase similarity among the original and fully transformed (i.e., passed through forward and backward cycles) features. Starting from the assumption that the PDFs could be observed as Gaussians embedded into the cone of the symmetric positive definite (SPD) matrices, various statistics-based as well as geodesic-ground-distance-based measures can be utilised as a part of the objective function within the training procedure.

Both BTS SLL CycleGAN and FMR CycleGAN models could be employed independently for a multitude of domain-specific style adaptation tasks, such as image-to-image translation or speech enhancement analysed in this paper, the first one requiring only a small number of the aligned data samples during initial training epochs. Baseline CycleGAN architecture is briefly described in Section 2. BTS SSL CycleGAN is presented in Section 3, and FMR CycleGAN in Section 4. In Section 5, experimental results are presented for a variety of image-to-image translation tasks and a speech enhancement, i.e., noise reduction task. Conclusions are drawn in Section 6.

2. CYCLEGAN (BASELINE APPROACH)

CycleGAN [9] is an unsupervised training approach aiming to learn a translation of samples from the originating domain X into a target domain Y. The architecture consists of two generators, $G_{X\to Y}$ (direct) and $G_{Y\to X}$ (inverse), and the associated adversarial discriminators, D_X and D_Y. The idea behind the adversarial loss is relatively simple. The generators $G_{X\to Y}$ and $G_{Y\to X}$, stimulated by discriminators D_X and D_Y, attempt to minimise the difference between $G_{X\to Y}(x)$ and y, as well as $G_{Y\to X}(y)$ and x, $x\in X$, $y\in Y$, i.e., $G_{X\to Y}(x)$ should be as close as possible to y and $G_{Y\to X}(y)$ should be as close as possible to x, while discriminators D_X and D_Y try to distinguish between real (x,y) and translated samples $(G_{Y\to X}(y),G_{X\to Y}(x))$.

The adversarial objectives (1) and (2)

$$\mathcal{L}_{adv}(G_{X\to Y},D_Y)=\mathbb{E}_{y\sim p_Y(y)}[\ln D_Y(y)] + \mathbb{E}_{x\sim p_X(x)}\left[\ln\left(1-D_Y\left(G_{X\to Y}(x)\right)\right)\right] \quad (1)$$

$$\mathcal{L}_{adv}(G_{Y\to X},D_X)=\mathbb{E}_{x\sim p_X(x)}[\ln D_X(x)] + \mathbb{E}_{y\sim p_Y(y)}\left[\ln\left(1-D_X\left(G_{Y\to X}(y)\right)\right)\right] \quad (2)$$

Equation 1 – Adversarial objectives.

where $p_X(x)$ and $p_Y(y)$ represent source and target data distributions, are additionally coupled with forward and backward cycle-consistency objectives, given by

$$\mathcal{L}_{cyc}(G_{X\to Y},G_{Y\to X})=\mathbb{E}_{x\sim p_X(x)}\left[\left\|G_{Y\to X}\left(G_{X\to Y}(x)\right)-x\right\|\right] + \mathbb{E}_{y\sim p_Y(y)}\left[\left\|G_{X\to Y}\left(G_{Y\to X}(y)\right)-y\right\|\right]$$

Equation 2 – Cycle-consistency objective.

providing cycle-consistent forward and backward mappings (i.e., after one full cycle, translated samples should be as close as possible to the original samples provided as inputs), and the identity loss

$$\mathcal{L}_{id}(G_{X\to Y},G_{Y\to X})=\mathbb{E}_{y\sim p_Y(y)}[\|G_{X\to Y}(y)-y\|] + \mathbb{E}_{x\sim p_X(x)}[\|G_{Y\to X}(x)-x\|]$$

Equation 3 – Identity objective.

regularizing the generators $G_{X\to Y}$ and $G_{Y\to X}$, producing near identity mappings in cases when real samples of the target domain are provided as inputs.

The optimisation problem can now be represented as

$$G_{X\to Y}{}^{*},G_{Y\to X}{}^{*}=\arg\min_{G_{X\to Y},G_{Y\to X}}\max_{D_X,D_Y}\mathcal{L}(G_{X\to Y},G_{Y\to X},D_X,D_Y)$$

Equation 4 – CycleGAN optimization problem.

where $\mathcal{L}(G_{X\to Y},G_{Y\to X},D_X,D_Y)$ represents a full CycleGAN objective function given by

$$\mathcal{L}(G_{X\to Y},G_{Y\to X},D_X,D_Y)= \mathcal{L}_{adv}(G_{X\to Y},D_Y)+\mathcal{L}_{adv}(G_{Y\to X},D_X)+ \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y},G_{Y\to X})+\lambda_{id}\mathcal{L}_{id}(G_{X\to Y},G_{Y\to X})$$

Equation 5 – CycleGAN objective function.

using λ_{cyc} and λ_{id} as the appropriate mixing coefficients.

3. BTS SSL CYCLEGAN

3.1. SEMI-SUPERVISED LEARNING

BTS SSL CycleGAN approach [1] represents a task-independent solution for unbalanced data domains, i.e., when one of the domains is fully observable, the other one is scarce, and a certain predefined number of data samples are matched, i.e., presented in pares containing the same core information (e.g., data structure, shape, object representation) for source and target data domains. For any predefined number of labelled (paired) data samples $\{(x_i,y_i)|i=1,\ldots,m\}\subset X\times Y$, a supervised training procedure is applied by introducing an additional $\|\cdot\|1$ norm term given in Equation 6 into the overall objective function given in Equation 5, enforcing similarity and closeness among the same-labelled data representations. The error is calculated for both direct and inverse mappings, averaged over pairs of correlated data samples.

If m is the number of correlated (paired) data samples, the SSL objective is given by

$$\mathcal{L}_{SSL}(G_{X\to Y},G_{Y\to X})=\frac{1}{m}\sum_{i=1}^{m}[\|G_{X\to Y}(x_i)-y_i\|_1 + \|G_{Y\to X}(y_i)-x_i\|_1]$$

Equation 6 – SSL objective.

meaning that the full BTS SLL CycleGAN objective function can now finally be defined as

$$\mathcal{L}(G_{X\to Y},G_{Y\to X},D_X,D_Y)= \mathcal{L}_{adv}(G_{X\to Y},D_Y)+\mathcal{L}_{adv}(G_{Y\to X},D_X) +\lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y},G_{Y\to X})+\lambda_{id}\mathcal{L}_{id}(G_{X\to Y},G_{Y\to X}) +\lambda_{SSL}\mathcal{L}_{SSL}(G_{X\to Y},G_{Y\to X})$$

Equation 7 – BTS SSL CycleGAN objective function.

for appropriate values of mixing coefficients λ_{cyc}, λ_{id}, and λ_{SSL}.

SSL strategy enables exploitation of an entire training dataset (for unlabelled samples, the standard CycleGAN objective function given in Equation 5 is applied instead of BTS SLL objective function given in Equation 7), which in turn prevents overfitting due to a limited number of paired data samples, providing better stability and increased accuracy.

3.2. BOOTSTRAPPING

Bootstrapping strategy is applied after a predefined number of training epochs, i.e., when the generator $G_{Y\to X}$ is sufficiently trained and reasonably reliable, to overcome the imbalance issue between the scarce domain X and the fully observable domain Y. Randomly generated samples, produced by generator $G_{Y\to X}$ previously initialised within the SLL training procedure described in Subsection 3.1 (or after the unsupervised procedure described in Section 2), are added to the training pool of discriminator of the scarce domain D_X (at the same time, training pool of the discriminator D_Y remains intact). The bootstrapping is periodically repeated in conjunction with the SLL strategy during subsequent training iterations (each time more precise), replacing (instead of accumulating) previously generated samples and increasing the number of training samples in the scarce domain X. Consequently, improved discrimination capabilities of the scarce domain discriminator D_X are obtained, eventually improving the general performance of the proposed BTS SLL CycleGAN algorithm, as proven by experiments.

4. FMR CYCLEGAN

Presuming the same internal structure of direct and inverse network generators $G_{X\to Y}$ and $G_{Y\to X}$, FMR CycleGAN approach [13] introduces an additional cycle-consistent loss calculated among the same-depth input-output feature-map tensors represented as PDFs. In the case of the direct CycleGAN generator $G_{X\to Y}$, $F^f_{X\to Y}(x) \in \mathbb{R}^{m^f\times n^f\times d}$ and $F^l_{X\to Y}(x) \in \mathbb{R}^{m^l\times n^l\times d}$ represent feature-map tensors of the first and the last $G_{X\to Y}$ layer calculated for sample $x\in X$. In the case of the inverse CycleGAN generator $G_{Y\to X}$, feature-map tensors of the first and the last layer of $G_{Y\to X}$ are denoted as $F^f_{Y\to X}(y) \in R^{m^f\times n^f\times d}$ and $F^l_{Y\to X}(y) \in R^{m^l\times n^l\times d}$ for sample $y\in Y$.

$F^{\{f,l\}}_{X\to Y}(x)$ and $F^{\{f,l\}}_{X\to Y}(x)$ can be reshaped into d-dimensional-column-based matrices of size $d\times(m^{\{f,l\}}\cdot n^{\{f,l\}})$ in the following way

$$F^{\{f,l\}}_{X\to Y\,mat}(x) = \left[F^{\{f,l\}}_{X\to Y\,1,1,:}(x)|\,...\,|F^{\{f,l\}}_{X\to Y\,m^{\{f,l\}},n^{\{f,l\}},:}(x)\right] \quad (1)$$
$$F^{\{f,l\}}_{Y\to X\,mat}(y) = \left[F^{\{f,l\}}_{Y\to X\,1,1,:}(y)|\,...\,|F^{\{f,l\}}_{Y\to X\,m^{\{f,l\}},n^{\{f,l\}},:}(y)\right] \quad (2)$$

Equation 8 – Feature map matrices.

Starting from the assumption that the underlying PDFs $f^{\{f,l\}}_{X\to Y}$ and $f^{\{f,l\}}_{Y\to X}$of feature maps $F^{\{f,l\}}_{X\to Y}$ and $F^{\{f,l\}}_{Y\to X}$ can be represented as d-dimensional multivariate Gaussians, their Maximum Likelihood (ML) estimates can be obtained as

$$\Sigma^{\{f,l\}}_{X\to Y}(x) = \frac{1}{m^{\{f,l\}}n^{\{f,l\}}}\sum_{i=1}^{m^{\{f,l\}}}\sum_{j=1}^{n^{\{f,l\}}}(F^{\{f,l\}}_{X\to Y\,i,j,:}(x) - \mu^{\{f,l\}}_{X\to Y}(x))(F^{\{f,l\}}_{X\to Y\,i,j,:}(x) - \mu^{\{f,l\}}_{X\to Y}(x))^T \quad (1)$$

$$\Sigma^{\{f,l\}}_{Y\to X}(y) = \frac{1}{m^{\{f,l\}}n^{\{f,l\}}}\sum_{i=1}^{m^{\{f,\ \}}}\sum_{j=1}^{n^{\{f,\ \}}}(F^{\{f,l\}}_{Y\to X\,i,j,:}(y) - \mu^{\{f,l\}}_{Y\to X}(y))(F^{\{f,l\}}_{Y\to X\,i,j,:}(y) - \mu^{\{f,l\}}_{Y\to X}(y))^T \quad (2)$$

Equation 9 – ML estimates of covariance matrices $\Sigma^{\{f,l\}}_{X\to Y}(x)$ and $\Sigma^{\{f,l\}}_{Y\to X}(y)$.

where

$$\mu^{\{f,l\}}_{X\to Y}(x) = \frac{1}{m^{\{f,l\}}n^{\{f,l\}}}\sum_{i=1}^{m^{\{f,l\}}}\sum_{j=1}^{n^{\{f,l\}}}F^{\{f,l\}}_{X\to Y\,i,j,:}(x) \quad (1)$$
$$\mu^{\{f,l\}}_{Y\to X}(y) = \frac{1}{m^{\{f,l\}}n^{\{f,l\}}}\sum_{i=1}^{m^{\{f,l\}}}\sum_{j=1}^{n^{\{f,l\}}}F^{\{f,l\}}_{Y\to X\,i,j,:}(y) \quad (2)$$

Equation 10 – ML estimates of mean vectors $\mu^{\{f,l\}}_{X\to Y}(x)$ and $\mu^{\{f,l\}}_{Y\to X}(y)$.

for any given $x\in X$ and $y\in Y$.

The proposed feature-map-based cycle-consistent loss term can now be defined as

$$\begin{aligned}&\mathcal{L}_{FMR}(G_{X\to Y}, G_{Y\to X}) = \\ &E_{x\sim p_X(x)}\left[d_{grd}\left(f^f_{X\to Y}(x), f^l_{Y\to X}\left(G_{X\to Y}(x)\right)\right)\right] + \\ &E_{y\sim p_Y(y)}\left[d_{grd}\left(f^f_{Y\to X}(y), f^l_{X\to Y}\left(G_{Y\to X}(y)\right)\right)\right]\end{aligned}$$

Equation 11 – FMR objective.

where d_{grd} represents some of the ground-based distances discussed in Section 5, providing the full FMR CycleGAN objective function as

$$\begin{aligned}&\mathcal{L}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y)= \\ &\mathcal{L}_{adv}(G_{X\to Y}, D_Y) + \mathcal{L}_{adv}(G_{Y\to X}, D_X) \\ &\qquad + \lambda_{cyc}\mathcal{L}_{cyc}(G_{X\to Y}, G_{Y\to X}) \\ &\qquad + \lambda_{id}\mathcal{L}_{id}(G_{X\to Y}, G_{Y\to X}) \\ &\qquad + \lambda_{FMR}\mathcal{L}_{FMR}(G_{X\to Y}, G_{Y\to X})\end{aligned}$$

Equation 12 – FMR CycleGAN objective function.

for given λ_{cyc}, λ_{id}, and λ_{FMR} as mixing coefficients.

5. EXPERIMENTAL RESULTS

5.1. IMAGE-TO-IMAGE TRANSLATION

In Table 1 and Table 2, the results are presented for the Pix2Pix network architecture presented in [5], baseline CycleGAN architecture proposed in [9], BTS SLL CycleGAN architecture described in Section 3 (including the results obtained using separate SSL and BTS training mechanisms referred in Subsections 3.1 and 3.2, respectively), and the FMR CycleGAN training strategy proposed in Section 4, in terms of the standard objective Peak Signal-to-Noise Ratio (PSNR) and Structural

Similarity Index (SSIM) measurements [1]. PSNR measurement is used as an energy-preserving measurement to estimate the quality of generated images versus their original counterparts. SSIM, on the other hand, evaluates image quality degradation as a perceived change in structural information. Both BTS SLL CycleGAN and FMR CycleGAN training architectures are built upon the baseline CycleGAN architecture, adding the proper loss term into the objective function, and the previously described bootstrapping mechanism in the case of the BTS SSL CycleGAN approach.

Three different image-to-image translation tasks have been conducted using various datasets (the Google Maps dataset, containing 1096 training images, the CityScapes dataset containing 2975 training images, and the CMP Facade dataset, containing 400 training images). The final scores were calculated using 50 generated images after 200 training epochs (a fixed learning rate value of 0.0002 was used for the first 100 epochs, decaying to zero during subsequent epochs).

The parameters λ_{cyc}, λ_{id}, λ_{SSL} and λ_{FMR} have all been fixed and set to 10. In the case of the FMR CycleGAN, various ground-based distances have been examined, namely, the robust L1-based distance (FMR$_{L1}$), the Kullback-Leibler divergence (FMR$_{KL}$), and the Log-Euclidean metric (FMR$_{LE}$) [13]. In the case of (semi-) supervised methods (Pix2Pix, SSL and BTS SSL), the size of the scarce domain has been manipulated {25, 50, 100}, changing the percentage of paired domain-to-domain training data. However, for the unsupervised training procedures (baseline CycleGAN, FMR$_{L1}$, FMR$_{KL}$, and FMR$_{LE}$ CycleGANs), the whole training dataset has been employed.

Both the semi-supervised learning and the bootstrapping training strategies contribute to the increase of average PSNR and SSIM values, simultaneously improving performances of the proposed BTS SSL CycleGAN algorithm in comparison with the baseline CycleGAN algorithm, and in some cases, even the fully-supervised Pix2Pix algorithm has been outperformed.

Table 1 - PSNR measures.

Dataset	S_X [%]	Pix2Pix	CycleGAN	SSL	BTS	BTS SSL	FMR$_{L1}$	FMR$_{KL}$	FMR$_{LE}$
CityScapes	25	19.98		18.30	17.30	18.77			
	50	20.45		18.95	17.18	19.04			
	100	19.51	17.12	20.03	17.75	20.47	17.89	18.86	17.34
CMP Facade	25	13.78		11.78	10.81	11.83			
	50	14.24		13.22	11.92	13.75			
	100	14.25	10.98	12.88	11.52	13.21	10.81	11.45	11.37
Google Maps	25	30.35		30.62	30.55	31.20			
	50	30.55		30.68	29.81	30.88			
	100	30.01	30.24	30.92	30.27	31.23	31.15	30.85	30.32

Table 2 – SSIM measures.

Dataset	S_X [%]	Pix2Pix	CycleGAN	SSL	BTS	BTS SSL	FMR$_{L1}$	FMR$_{KL}$	FMR$_{LE}$
CityScapes	25	0.60		0.59	0.58	0.61			
	50	0.64		0.61	0.59	0.64			
	100	0.59	0.54	0.58	0.63	0.65	0.58	0.65	0.56
CMP Facade	25	0.35		0.31	0.27	0.32			
	50	0.40		0.37	0.28	0.40			
	100	0.42	0.27	0.33	0.35	0.41	0.31	0.29	0.28
Google Maps	25	0.67		0.73	0.75	0.77			
	50	0.68		0.75	0.76	0.79			
	100	0.69	0.73	0.75	0.77	0.81	0.76	0.74	0.73

Also, due to the additional alignment between feature maps of input-output generators layers, compared with the baseline CycleGAN algorithm, FMR CycleGAN provides better results in most cases and for all ground distances used (FMR_{L1}, FMR_{KL}, and FMR_{LE}). Visually pleasing and structurally more accurate results have been obtained using the proposed BTS SSL and FMR CycleGAN algorithms in comparison to the baseline CycleGAN and Pix2Pix (Figure 1).

5.2. SPEECH ENHANCEMENT

The BTS SSL CycleGAN algorithm described in this paper has also been applied within a speech enhancement, i.e. noisy to clean speech translation (style adaptation) task [2]. The results are given in Table 3 and Table 4, in terms of the Perceptual Evaluation of Speech Quality (PESQ, [14]) measures PEQMOS and MOSLQO, and the Virtual Speech Quality Objective Listener (ViSQOL, [15]) measures VISQOL and NSIM, respectively. The architecture presented in [16] has been employed, enabling parallel processing of the sequential data. 5000 training epochs have been conducted, using the generator(s) learning rate of 0.0002, and the discriminator(s) learning rate of 0.0001.

Spectral features have been extracted from randomly chosen signal snippets of 128 frames. The training database contains 200 clean and 200 artificially generated noisy speech samples (including both stationary and non-stationary noise components, such as background speech, traffic noise, creaking, etc.). The results are averaged over 50 test samples, by comparisons between the transformed noisy to clean speech samples and their clean speech counterparts. Presented results support our previous observations (BTS and SSL components separately, as well as in conjunction, provide more favourable results compared to the baseline CycleGAN algorithm for any percentage of the scarce noisy speech domain used in a supervised manner). Figure 2 shows the spectrograms of a selected noisy part of a noisy speech signal transformed using the proposed algorithms (blue-green-yellow colour range symbolises lowest to highest noise energy). While preserving the vocal component in most cases (speech around non-stationary noise components has been filtered in some cases), noise has been significantly reduced, which also corresponds to the results obtained by subjective (listening) evaluations.

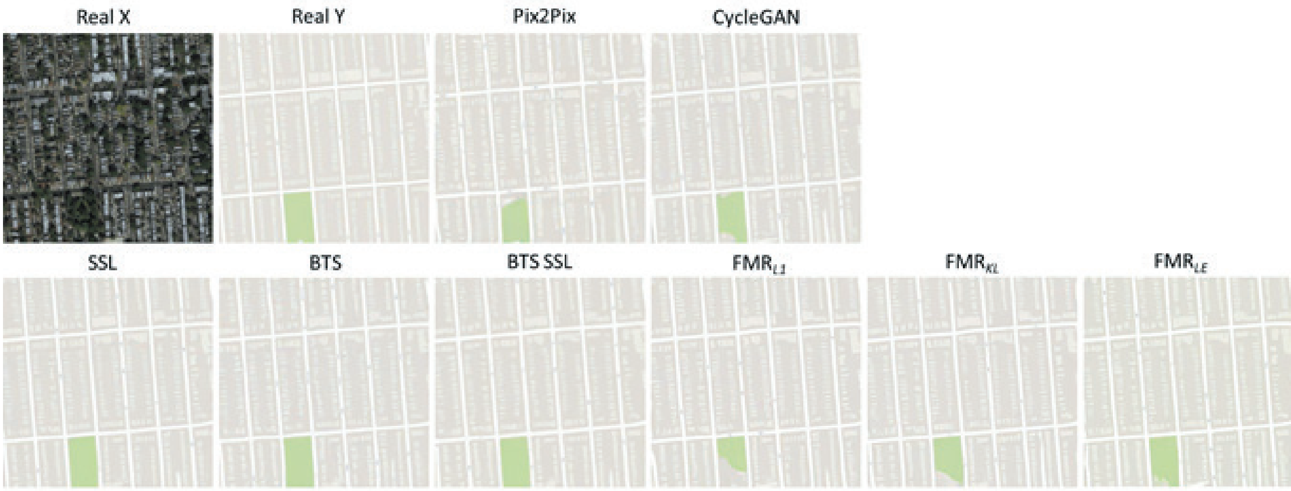

Figure 1 – Google Maps image-to-image translation task (100% of the scarce domain used).

Table 3 - PESQ measures.

S_X [%]	CycleGAN		SSL		BTS		BTS SSL	
	PEQMOS	MOSLQO	PEQMOS	MOSLQO	PEQMOS	MOSLQO	PEQMOS	MOSLQO
25			0.831	1.179	0.832	1.171	0.837	1.178
50			0.864	1.184	0.842	1.186	0.857	1.191
100	0.828	1.178	0.853	1.211	0.850	1.232	0.867	1.238

Table 4 – ViSQOL measures.

S_x [%]	CycleGAN		SSL		BTS		BTS SSL	
	VISQOL	NSIM	VISQOL	NSIM	VISQOL	NSIM	VISQOL	NSIM
25			1.384	0.572	1.399	0.573	1.392	0.579
50			1.409	0.572	1.410	0.575	1.425	0.581
100	1.369	0.569	1.436	0.576	1.391	0.575	1.452	0.585

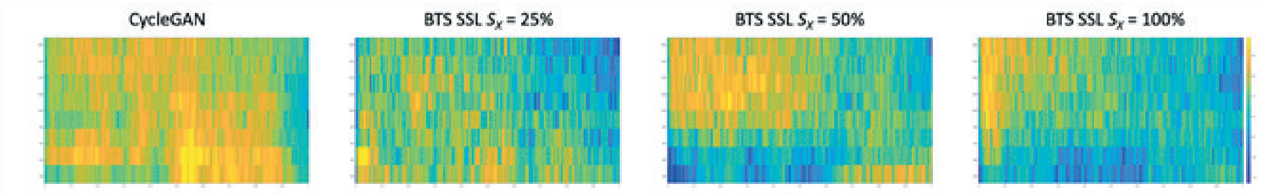
CycleGAN
BTS SSL S_x = 25%
BTS SSL S_x = 50%
BTS SSL S_x = 100%

Figure 2 - Spectrograms of the transformed noisy speech signal (noise component).

6. CONCLUSION

In this paper, the performances of the proposed BTS SSL CycleGAN algorithm, introducing a semi-supervised learning strategy and a bootstrapping method, and the FMR CycleGAN algorithm, adding an additional feature map regularisation, have been compared among each other and also against the baseline unsupervised CycleGAN and supervised Pix2Pix approaches. The first one improves performance in the case of highly imbalanced domain-to-domain style adaptation tasks. The second one achieves more favourable results in an unsupervised training scenario, compared to the baseline unsupervised CycleGAN approach, and close to the supervised Pix2Pix approach. Improvements behind the bootstrapping logic of the BTS SSL, reducing computational complexity of geodesic and information distances calculations during the FMR training phase, improving the performance of speech enhancement around non-stationary noise components, and analysing additional use case scenarios, such as speech style (emotion) transformation, will be the subject of future study.

7. ACKNOWLEDGEMENTS

This research was supported by the Science Fund of the Republic of Serbia, #6524560, AI-S-ADAPT, and by the Serbian Ministry of Education, Science and Technological Development through the project no. 45103–68/2020–14/200156: ”Innovative Scientific and Artistic Research from the Faculty of Technical Sciences Activity Domain”.

8. REFERENCES

[1] L. Krstanović, B. Popović, M. Janev and B. Brkljač, “Bootstrapped SSL CycleGAN for Asymmetric Domain Transfer,” *Applied Sciences*, vol. 12, no. 7, p. 3411, 2022.

[2] B. Popović, L. Krstanović, M. Janev, S. Suzić, T. Nosek and J. Galić, “Speech Enhancement Using Augmented SSL CycleGAN,” in 30th *European Signal Processing Conference (EUSIPCO)*, 2022.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, “Generative Adversarial Networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139-144, 2020.

[4] O. Ronneberger, P. Fischer and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in 18th *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[5] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, “Image-to-Image Translation With Conditional Adversarial Networks,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] C. Wang, H. Zheng, Z. Yu, Z. Zheng, Z. Gu and B. Zheng, “Discriminative Region Proposal Adversarial Networks for High-Quality Image-to-Image Translation,” in *European Conference on Computer Vision (ECCV)*, 2018.

[7] T. R. Shaham, M. Gharbi, R. Zhang, E. Shechtman and T. Michaeli, “Spatially-Adaptive Pixelwise Networks for Fast Image Translation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[8] T. Park, M.-Y. Liu, T.-C. Wang and J.-Y. Zhu, "Semantic Image Synthesis With Spatially-Adaptive Normalization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[9] J.-Y. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] M.-Y. Liu and O. Tuzel, "Coupled Generative Adversarial Networks," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[11] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim, "Learning To Discover Cross-Domain Relations With Generative Adversarial Networks," in *International Conference on Machine Learning*, 2017.

[12] M.-Y. Liu, T. Breuel and J. Kautz, "Unsupervised Image-to-Image Translation Networks," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] L. Krstanović, B. Popović, M. Janev and B. Brkljač, "Feature Map Regularized CycleGAN for Domain Transfer," *Mathematics*, vol. 11, no. 2, p. 372, 2023.

[14] A. W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) – A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[15] A. Hines, J. Skoglund, A. Kokaram and N. Harte, "ViSQOL: The Virtual Speech Quality Objective Listener," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2012.

[16] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks," in *26th European Signal Processing Conference (EUSIPCO)*, 2018.