




Image-to-Image Translation with CNN Based Perceptual Similarity Metrics

CNN Tabanlı Algısal Benzerlik Metrikleriyle Görüntüden Görüntüye Dönüşüm

Sara ALTUN GÜVEN^{*1} , Emrullah ŞAHİN² , M. Fatih TALU³ 

¹Bilgisayar Mühendisliği Bölümü, Tarsus Üniversitesi, Mersin, Türkiye

²Yazılım Mühendisliği Bölümü, DEFG Üniversitesi, Kütahya, Türkiye

³Bilgisayar Mühendisliği Bölümü, DEFG Üniversitesi, Malatya, Türkiye

(saraguen@tarsus.edu.tr, emrullah.sahin@dpu.edu.tr, fatihtalu@gmail.com)

Received:Jan.31,2024

Accepted:Mar.13,2024

Published:Jun.01,2024

Özetçe— Görüntüden görüntüye çeviri, farklı alanlardaki görüntüleri dönüştürme sürecidir. Çekişmeli Üretici Ağlar (Generative Adversarial Networks - GANs) ve Evrimsel Sinir Ağları (Convolutional Neural Networks - CNNs), görüntü çevirisinde yaygın olarak kullanılan tekniklerdir. Bu çalışma, GAN mimarileri için etkili bir kayıp fonksiyonunu bulmayı ve daha kaliteli görüntüler üretmeyi hedeflemektedir. Bu amaçla, temel bir GAN mimarisi olan Pix2Pix yöntemindeki kayıp fonksiyonları üzerinde deneysel çalışmalar yapılmıştır. Pix2Pix yönteminde kullanılan mevcut kayıp fonksiyonu Mean Absolute Error (MAE) olarak bilinen \mathcal{L}_1 metriğidir. Bu çalışmada, Pix2Pix mimarisinde kayıp fonksiyonuna konvolüsyon tabanlı algısal benzerlik metriklerinin (CONTENT, LPIPS ve DISTs) etkileri incelenmiştir. Ayrıca, görüntüden görüntüye çevirme üzerindeki etkiler, orijinal \mathcal{L}_1 kaybıyla birlikte $\mathcal{L}_1_CONTENT$, \mathcal{L}_1_LPIPS ve \mathcal{L}_1_DISTs algısal benzerlik metrikleri kullanılarak yüzde 50 oranında analiz edildi. Yöntemlerin performans analizleri çeşitli açık erişimli veri setleri üzerinde gerçekleştirilmiştir. Görsel sonuçlar, geleneksel (FSIM, HaarPSI, MS-SSIM, PSNR, SSIM, VIFp ve VSI) ve güncel (FID ve KID) görüntü karşılaştırma metrikleri ile analiz edildi. Sonuç olarak, GAN mimarilerinin kayıp fonksiyonu için konvansiyonel yöntemler yerine konvolüsyon tabanlı yöntemler kullanıldığında daha iyi sonuçlar elde edildiği gözlemlendi. Ayrıca, LPIPS ve DISTs yöntemlerinin gelecekte GAN mimarilerinin kayıp fonksiyonunda kullanılabileceğine dair umut verici sonuçlar alınmıştır.

Anahtar Kelimeler : Derin öğrenme, Benzerlik metrikleri, Görüntüden görüntüye dönüşüm, Evrimsel sinir ağ, Çekişmeli üretici ağ

Abstract— Image-to-image translation is the process of transforming images from different domains. Generative Adversarial Networks (GANs), and Convolutional Neural Networks (CNNs) are widely used in image translation. This study aims to find the most effective loss function for GAN architectures and synthesize better images. For this, experimental results were obtained by changing the loss functions on the Pix2Pix method, one of the basic GAN architectures. The exist loss function used in the Pix2Pix method is the Mean Absolute Error (MAE). It is called the \mathcal{L}_1 metric. In this study, the effect of convolutional-based perceptual similarity CONTENT, LPIPS, and DISTs metrics on image-to-image translation was applied on the loss function in Pix2Pix architecture. In addition, the effects on image-to-image translation were analyzed using perceptual similarity metrics ($\mathcal{L}_1_CONTENT$, \mathcal{L}_1_LPIPS , and \mathcal{L}_1_DISTs) with the original \mathcal{L}_1 loss at a rate of 50%. Performance analyzes of the methods were performed with the Cityscapes, Denim2Mustache, Maps, and Papsmeat datasets. Visual results were analyzed with conventional (FSIM, HaarPSI, MS-SSIM, PSNR, SSIM, VIFp and VSI) and up-to-date (FID and KID) image comparison metrics. As a result, it has been observed that better results are obtained when convolutional-based methods are used instead of conventional methods for the loss function of GAN architectures. It has been observed that LPIPS and DISTs methods can be used in the loss function of GAN architectures in the future.

Keywords : Deep learning, Similarity metrics, Image to image translation, Convolutional neural network, Generative adversarial networks

1. Introduction

Deep learning-based studies have been advancing rapidly in recent years. One of the evolving methods in this field is image synthesis. Image synthesis involves the process of editing, manipulating, translating an image, or generating an image from a signal. Convolutional Neural Networks (CNNs) (Zhu et al., 2017) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are extensively utilized in research within this domain. CNN-based studies in feature extraction processes, such as image synthesis or pattern recognition, surpass conventional methods (Karpathy et al., 2014; Koushik et al., 2016; Guo et al., 2016). One of the approaches to image synthesis using CNN is PixelRNN, which was proposed by Oord et al. (June 2016). PixelRNN is a deep neural network designed to predict output images from input images in the spatial domain, encoding the full set of dependencies in the pattern by modeling the discrete probability of raw pixel values (Oord et al., June 2016). The dataset used for experimentation is ImageNet, and the results demonstrate the consistency of the method. Oord et al. (2016) also introduced another generative model, PixelCNN. PixelCNN employs autoregressive links during image synthesis and is faster than PixelRNNs in the training phase. The proposed method proves advantageous as dataset patterns increase in complexity during image synthesis. Salimans et al. (2017) proposed the PixelCNN++ method. PixelCNN++ is a generative model similar to PixelCNN but simplifies its structure, conditioning all pixels instead of R/G/B subpixels. It also incorporates Dropout regularization to regularize the model. Chen et al. (2017) presented an image synthesis approach called SCA-CNN. SCA-CNN combines spatial information and image color channel information in the image, outperforming current attention-based image synthesis methods.

Along with the high performance in CNN-based studies, Generative Adversarial Networks (GANs) also exhibit superior performance in image synthesis processes (Liu et al., 2017; Liu et al., 2016; Kingma et al., 2013; Wang et al., 2018). GANs, a method developed based on deep learning, were proposed by Ian Goodfellow et al. (Goodfellow et al., 2014) in 2014. This method comprises two neural networks operating in contention. Numerous studies have been conducted on image synthesis with GAN architectures. For instance, Liu et al. (2017) introduced the UNIT method, which performs unsupervised image-to-image translation based on GANs (Liu et al., 2017). UNIT combines the CoGAN (Liu et al., 2016) and VAE (Kingma et al., 2013) methods to achieve unsupervised image-to-image translation. The UNIT method involves six networks: two encoders, generators, and discriminators. Input and output images must have similar areas for optimal performance in this model. Wang et al. (2018) proposed the Pix2PixHD architecture to address the synthesis problem (Wang et al., 2018). This architecture is based on a generator network and three scaled discriminators. Output images have dimensions of 2048×1024 . Liu et al. (2020) proposed a model that synthesizes representation content by separating it from domain attributes. Named GMM-UNIT, this model uses the Gaussian mixture model (GMM) for the hidden field attribute. GMM-UNIT has two main advantages: it allows translation in multiple fields and enables interpolation between fields and extrapolation within invisible fields. Royer et al. (2020) introduced the XGAN model based on the loss of semantic consistency. This model is a binary contention autoencoder capturing the shared feature representation of both areas to learn standard feature-level information rather than pixel-level information. It utilizes the loss of semantic consistency in both domains to preserve the image's semantic content across domains. In 2018, Frid-Adar et al. created synthetic medical images using GAN in image synthesis. It was observed that the produced images can be used in data augmentation and medical image classification, thereby improving the performance of CNN. Today, various GAN architectures are employed in multiple areas, as highlighted by Isola et al. (2017) and Zhu et al. (2017).

One of the major challenges in image-to-image translation lies in the insufficient evaluation of the similarity between the original and the generated image. As a result, multiple studies have been conducted to address the issue of assessing image quality, which falls under the domain of Image Quality Assessment (IQA). IQA aims to measure various aspects of image quality, including structural, textural, diversity, and signal strength. LPIPS (Zhang et al., 2018), DISTs (Ding et al., 2020), and CONTENT (Gatys et al., 2015) methods can be employed both within IQA frameworks and as loss functions during image synthesis. Given that these three methods are based on convolutional neural networks, they are examined within a CNN context in this study. For the comparison of image synthesis outputs, traditional methods such as FSIM (Zhang et al., 2011), HaarPSI (Reisenhofer et al., 2018), PSNR (Fardo et al., 2016), MS-SSIM (Wang et al., 2003), SSIM (Wang et al., 2004), VIFp (Sheikh and Bovik, 2006), and VSI (Zhang et al., 2014), as well as modern FID (Heusel et al., 2003) and KID (Bin'kowski et al., 2003) IQA methods, were utilized. Numerous studies in the literature explore IQA methods and leverage them in image analysis (Heusel et al., 2003; Bin'kowski et al., 2003; Choi et al., 2020; Ding et al., 2021; Sim et al., 2020; Borasinski et al., 2022; Peng et al., 2022).

This study examined the impact of altering the loss function in a fundamental MMS architecture on image synthesis. The architecture is based on the Pix2Pix method (Isola et al., 2017), which utilizes supervised learning techniques introduced by Isola et al. in 2017. In supervised synthesis, the loss is calculated as the distance between the estimated output (y) generated from the input and the real image (x). This loss value is then utilized to update both the generator (G) and discriminator (D) networks. The original Pix2Pix method employs the mean absolute error (MAE) loss function. In this study, the influence of using LPIPS, DISTs, and CONTENT methods as loss

functions in GANs was investigated. These methods are CNN-based and can serve as measures of similarity between images. Additionally, these methods are founded on the VGG (Simonyan and Zisserman, 2014) network. The recently proposed LPIPS method encourages reverse mapping to learn while emphasizing perceptual similarity between fake and real images reconstructed by the generator network. It also gauges average feature distances between synthesized samples. A higher LPIPS score indicates greater diversity among rendered images. This method assesses structure and tissue similarity akin to SSIM. Suzuki et al. (2021) observed in their study that utilizing this method yielded results equivalent to visual outputs (Suzuki et al., 2021). Chuan et al. (2018) investigated a hybrid content similarity metric, using the CONTENT method as an example. The study analyzed four different datasets with various visual characteristics (Chuan et al., 2018).

The goal is to identify a continuous and highly accurate loss function. In the analysis of results, both traditional and contemporary image comparison metrics were employed. The image synthesis outcomes of the methods are presented in both visual and tabular formats based on these metrics. The primary contributions of favoring modern CNN-based methods over conventional loss functions can be summarized as follows:

- It has been demonstrated that it can be utilized as a general loss function in GAN methods.
- It has been observed to positively influence the results of image synthesis.
- It has been investigated and found that it leads to better synthesis of textural structures in the image.

The study’s most significant contribution to the literature has been the determination of the loss function and similarity metric, which maximizes the accuracy of the image synthesis process made with the GAN approach. It is predicted that the ease of use instead of most loss functions will guide most research in the future. The remainder of the manuscript follows; details of materials and methods explained in Section 2. Experimental results explained in Section 3. Finally, the conclusion is given in Section 4.

2. Materials and Methods

2.1. Pix2Pix

One of the extensively employed GAN architecture methods is Pix2Pix, based on DCGAN (Radford et al., 2015). This method comprises a generator and a discriminator network. The generator utilizes the U-Net architecture (Ronneberger et al., 2015), while the discriminator network employs PatchGAN (Li and Wand, 2016). The loss functions for the generator and discriminator in the Pix2Pix method are provided in Eq. (1, 2, 3).

$$L_D = \|D(X, Y) - 1\|_2 + \|D(X, G(X))\|_2 \quad (1)$$

$$L_G = \|D(X, G(X)) - 1\|_2 \quad (2)$$

$$\mathcal{L}_{GAN}(G, D, X, Y) = L_G + L_D \quad (3)$$

It has been observed that the blur level is high in the images synthesized with Pix2Pix. Isola et al. (2017) added the \mathcal{L}_1 regularization term to the loss of the generator architecture to remove some fuzziness. The loss function of the updated Pix2Pix is as follows Eq. (4):

$$G^* = \underset{G}{\operatorname{argmin}} \underset{D}{\operatorname{max}} \mathcal{L}_{GAN}(G, D, X, Y) + \lambda \mathcal{L}_{\text{Reg}} \quad (4)$$

$$\mathcal{L}_{\text{Reg}} = \|Y - G(X)\|_1$$

2.2. CNN-Based Loss Function

There are three commonly used measures in the literature that employ trained CNN architectures to generate similarity metrics between pairs of images: CONTENT, LPIPS, and DISTs. While statistical methods focus on pixel values, convolutional methods concentrate on image content (Zhang et al., 2018; Ding et al., 2020; Gatys et al., 2015; Zhang et al., 2011). Therefore, in this study, the performance of CNN-based architectures in image synthesis was analyzed using these loss functions. These three CNN-based methods share similarities and utilize trained VGG networks. Although they have demonstrated significant success as image benchmarks, the high computational cost and lack of interpretability may potentially hinder their practical applicability (Ding et al., 2021). In this study, the VGG19 network was employed to calculate the loss for CONTENT, LPIPS, and DISTs. The VGG19 network consists of 16 convolutional layers and 5 pooling layers (Simonyan and Zisserman, 2014), organized into five blocks concluding with a pooling layer. Figure 1 illustrates which blocks the CNN-based loss functions utilize the outputs from the VGG19 network. Additionally, specific weights of the VGG19 network trained according to the datasets from the original studies of each method were employed.

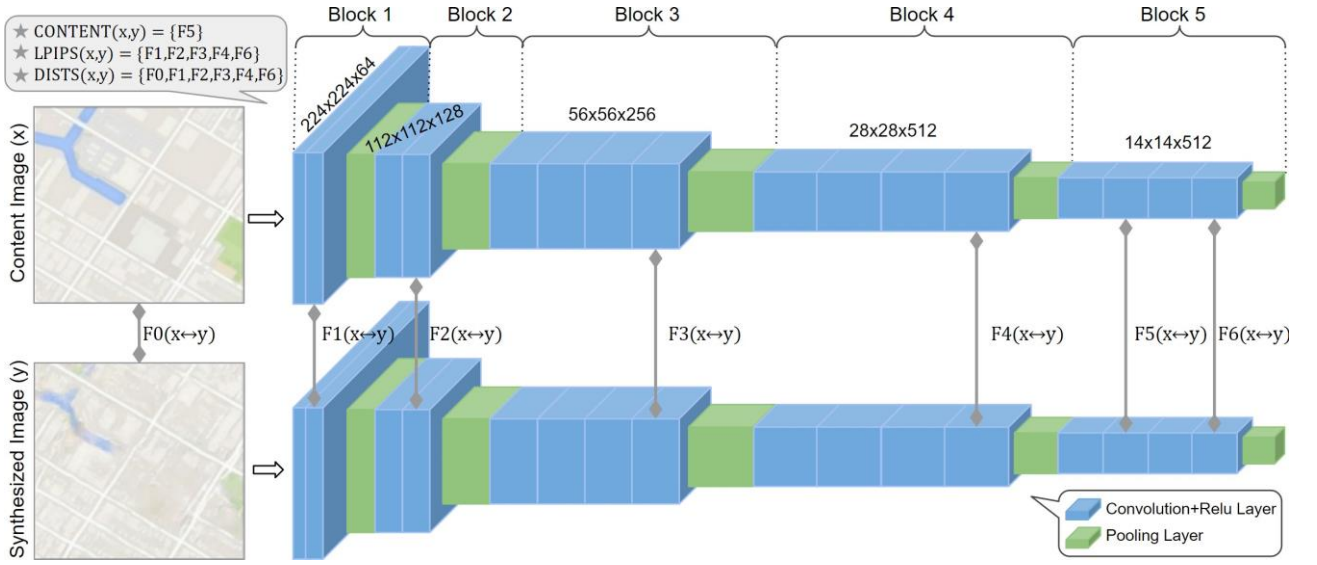


Figure 1. CNN-based CONTENT, LPIPS and DISTIS methods with VGG19 network.

2.2.1. CONTENT

The concept of CONTENT loss originated from the Neural Style Transfer study by Gatys et al. (2015). The value of this cost function is computed between the output of each block by providing both the content and target (synthesized) images to the VGG networks. The CONTENT cost aims to preserve the essential characteristics of the content display, as proposed by Gatys et al. (2015). This loss function is illustrated in Figure 1 on the VGG19 network. The similarity value is calculated between the outputs in the fifth block, denoted as F5, of the network as shown in the figure. After the outputs of the blocks pass through the Rectified Linear Unit (ReLU) activation function, they are subtracted from each other using the Mean Squared Error (MSE) method (Mihelich et al., 2020). The main formula of this metric is shown in Eq. (5). In the equation, the content image (x), the target image (y), and the list of block outputs (n) are expressed.

$$CONTENT(x, y) = 1 - \sum_{i=1}^n (x_i - y_i)^2 \quad (5)$$

2.2.2. LPIPS

LPIPS (Learned Perceptual Image Patch Similarity) is one of the most recent metrics used to measure perceptual similarity between pairs of images (Zhang et al., 2018). Simultaneously, this metric employs deep attributes that mimic human perception (Ding et al., 2021). Figure 1 illustrates the activation outputs on which the LPIPS function is computed using the VGG19 network. Accordingly, the outputs of same-level layers (F1-4, F6) for the x and y images in the VGG19 network are normalized and subtracted from each other. The resulting data is scaled, and the outputs are transformed into a single vector form (Zhang et al., 2018). The loss is obtained by calculating the vector norm using the Mean Squared Error (MSE) method.

The numerical representation of this method is given in Eq. (6). The w_i coefficient in the equation shows the perceptual importance of each layer.

$$LPIPS(x, y) = 1 - \sum_{i=1}^n w_i * (x_i - y_i)^2 \quad (6)$$

2.2.3. DISTIS

The DISTIS (Deep Image Structure and Texture Similarity) function uses VGG network (Ding et al., 2020). This method uses l2 pooling instead of general max pooling in the VGG network. The DISTIS function consists of

a combination of structure and texture similarity. It resists slight geometric distortions and performs well on textural images (Ding et al., 2020; Ding et al., 2021). DISTS distance between the x and y input images over the block outputs (F0-4, F6) in Figure 1. The main formula of this metric is shown in Eq. (7). The coefficients (α) and (β) in the equation are previously trained particular values from the original run of the DISTS function. In equation include the structural(s) and textural(t) functions.

$$DISTS(x, y) = 1 - \sum_{i=1}^n \alpha_i * s(x_i - y_i) + \beta_i * t(x_i - y_i)$$

$$s(x_i - y_i) = \frac{2 * \text{mean}(x_i) * \text{mean}(y_i) + c_1}{\text{mean}(x_i)^2 + \text{mean}(y_i)^2 + c_1}$$

$$t(x_i - y_i) = \frac{2 * \text{cov}(x_i, y_i) + c_2}{\text{var}(x_i) + \text{var}(y_i) + c_2}$$
(7)

2.2.4. Use of metrics with \mathcal{L}_1

While calculating the smoothing term, the effects on image-to-image translation were analyzed using perceptual similarity metrics ($\mathcal{L}_1_CONTENT$, \mathcal{L}_1_LPIPS , and \mathcal{L}_1_DISTS) with the original \mathcal{L}_1 loss at a rate of 50%. The generator network loss of Pix2Pix is updated as follows Eq. (6).

$$G^* = \underset{G}{\operatorname{argmin}} \underset{D}{\operatorname{max}} \mathcal{L}_{GAN}(G, D, X, Y) + \lambda \{0.5 * \mathcal{L}_1 + 0.5 * \mathcal{L}_{\{LPIPS, DISTS, CONTENT\}}(Y, G(X))\}$$
(8)

2.3. Image Comparison Metrics

2.3.1. Conventional methods

In this section, brief definitions of conventional image comparison metrics are provided. More detailed information can be obtained from the cited resources if desired.

Feature-Based Similarity Index Measurement (FSIM): Compares the phase coherence and gradient magnitude properties of image pairs (Zhang et al., 2011).

Haar Wavelet-Based Perceptual Similarity Index (HaarPSI): Based on comparing local wavelet coefficients extracted from image patches (Reisenhofer et al., 2018).

Structural Similarity Index Metric (SSIM): Utilizes simple statistical moments, such as mean (μ) and standard deviation (σ), to determine the similarity score of image pairs (Wang et al., 2004).

Multi-Structural Similarity Index Metric (MS-SSIM): Combines SSIM results calculated at different resolution levels (Wang et al., 2003).

Peak Signal Noise Ratio (PSNR): Widely used objective image signal quality metric. However, PSNR values may not correlate well with perceived image quality due to the complex, highly nonlinear nature of the human visual system (Fardo et al., 2016).

Visual Information Fidelity (VIF): Uses natural scene statistical models (NSS) with a distortion model to measure information shared between fake and original images (Sheikh and Bovik, 2006).

Visual Saliency-Induced Index (VSI): Assumes that a disturbance in one area that attracts the observer's attention is more disturbing than in another area. It aims to weigh local distortions with a local clarity map (Zhang et al., 2014).

In this section, brief definitions of conventional image comparison metrics are given. If desired, more comprehensive information can be obtained from the given resources. Feature-Based Similarity Index Measurement (FSIM) compares image pairs' phase coherence and gradient magnitude properties (Zhang et al., 2011). Haar Wavelet-Based Perceptual Similarity Index (HaarPSI) is based on comparing local wavelet coefficients extracted from image patches (Reisenhofer et al., 2018). In the Structural Similarity Index Metric (SSIM), a few simple statistical moments, such as mean (μ) and standard deviation (σ) are used to obtain the similarity score of the image pairs (Wang et al., 2004). In the Multi-Structural Similarity Index Metric (MS-SSIM),

SSIM results calculated at different resolution levels are combined (Wang et al., 2003). Peak Signal Noise Ratio (PSNR) is the most widely used objective image signal quality metric.

However, the PSNR values do not correlate well with perceived image quality due to the complex, highly nonlinear nature of the human visual system (Fardo et al., 2016). Visual Information Fidelity (VIF) utilizes natural scene statistical models (NSS) along with a distortion model to measure information shared between fake and original images (Sheikh and Bovik, 2006). The Visual Saliency-Induced Index (VSI) assumes that a disturbance in one area that attracts the observer’s attention is more disturbing than in another area. It attempts to weigh local distortions using a local clarity map (Zhang et al., 2014).

2.3.2. Up-to-date methods

In this section, brief definitions of contemporary image comparison metrics are provided.

Fréchet Inception Distance (FID): A performance measure used to assess the quality of images generated from GANs. FID compares the distribution of generated and real images (Heusel et al., 2017).

Kernel-Inception Distance (KID): A metric similar to FID that utilizes the squared Maximum Average Error (MAE) between images. It offers an advantage over the FID metric as it does not assume a parametric form in the distribution of activations and incorporates the ReLU activation function (Bin’kowski et al., 2018).

2.4. Datasets

Image synthesis analysis of the original and updated architectures was conducted using four different datasets.

Cityscapes Dataset: This dataset comprises video images of varying lengths captured from city streets (Cordts et al., 2016). It is commonly employed for evaluating the performance of partitioning algorithms. The dataset labels objects such as cars, roads, lanes, and traffic lights.

Denim2Mustache Dataset: This dataset contains 950 image pairs (denim-mustache) (Şahin and Talu, 2021). The images include front and back photos of three different denim products (trousers, skirts, and shorts), with mustache mask images overlaid on them. Each image has a size of $256 \times 256 \times 3$.

Maps Dataset: The Maps dataset (Isola et al., 2017) is a large collection of real-world images obtained from Google Maps. It includes labels for objects such as roads, parking areas, and grass areas in these images.

Papsmear Dataset: The Papsmear dataset (Altun and Talu, 2022) consists of 450 Papsmear-Mask image pairs. The images in the dataset depict cytoplasm, nucleus, white blood cell, bacillus, and speckle objects along with masks for these objects. The image dimensions are $256 \times 256 \times 3$.

3. Results

In order to understand the effect of the Pix2Pix method when used with different loss functions, the images in the Cityscapes dataset were used first. The translation results from this dataset are shown in Figure 2. The similarity values obtained as a result of the process are shown in Table 1. When the results of conventional similarity metrics were examined, it was observed that the LPIPS loss function was successful in MS-SSIM, SSIM, and VIFp similarity metrics. It is seen that the \mathcal{L}_1 -CONTENT function provides high success in FSIM, HaarPSI, PSNR, and VSI similarity metrics. These results mean high-fidelity translation can be made in Cityscapes images using LPIPS and \mathcal{L}_1 -CONTENT functions. When the FID and KID results from the up-to-date metrics in the Cityscapes dataset are examined, it is seen that the translation result of the DISTS function is adequate.

The second dataset analyzed in the article is Denim2Mustache. It is seen that the regression process is performed with the Denim2Mustache dataset while the classification is made in the Cityscapes, Maps, and Papsmear datasets. The performances of the proposed translation methods in both classification and regression processes are evaluated. The Denim2Mustache dataset contains only images of jeans objects; the output is a grayscale image (Mustache motifs). The results obtained for this dataset are shown in Figure 3. Numerical similarity results are given in Table 1. When looking at the conventional similarity measurement metrics in the Denim2Mustache dataset, it has been observed that the CONTENT method gives better results than FSIM, HaarPSI, MS-SSIM, PSNR, SSIM, and VSI. For VIFp. It was observed that the LPIPS function gave better results than the other methods and is more efficient than the up-to-date FID and KID metrics in Table 1.

The third dataset analyzed in the article is Maps. One image of the Maps dataset contains objects such as roads, parks, rivers, and houses, while the other image contains these objects’ graphical (partitioned) form. The visual results of the Maps dataset are given in Figure 4. According to these images, the performance of the \mathcal{L}_1 -DISTS function is apparent. In Table 1, according to the conventional similarity metrics FSIM, HaarPSI, PSNR, SSIM,

VSI for the MAPS dataset, the \mathcal{L}_1 -LPIPS function has observed that the LPIPS function gives better results than MS-SSIM, VIFp. When the up-to-date FID and KID similarity metrics are examined in Table 1, the DISTs function was successful.

Table 1. Performance comparison of architectures in image synthesis. (Rows: CNN-based loss functions and datasets, Columns: conventional and up-to-date similarity metrics)

Similarity Metrics										
Dataset	Loss function	Conventional							Up-to-date	
		FSIM	HaarPSI	MS-SSIM	PSNR	SSIM	VIFp	VSI	FID	KID
Cityscapes	\mathcal{L}_1	0.631	0.344	0.414	15.53	0.387	0.033	0.860	5.41	0.053
	LPIPS	0.645	0.353	0.438	15.81	0.416	0.039	0.869	3.23	0.045
	\mathcal{L}_1 -LPIPS	0.642	0.352	0.435	15.77	0.408	0.038	0.866	3.19	0.044
	DISTS	0.630	0.338	0.399	15.38	0.389	0.032	0.862	7.77	0.036
	\mathcal{L}_1 -DISTS	0.636	0.346	0.422	15.82	0.399	0.035	0.866	9.08	0.038
	CONTENT	0.644	0.348	0.414	15.39	0.402	0.036	0.869	16.5	0.097
	\mathcal{L}_1 -CONTENT	0.649	0.359	0.430	16.04	0.409	0.038	0.872	107.6	0.086
D2M	\mathcal{L}_1	0.824	0.561	0.854	16.96	0.842	0.287	0.928	207.0	0.121
	LPIPS	0.824	0.559	0.855	16.87	0.846	0.303	0.927	30.7	0.042
	\mathcal{L}_1 -LPIPS	0.822	0.556	0.852	16.84	0.843	0.292	0.926	57.7	0.058
	DISTS	0.825	0.551	0.847	16.84	0.839	0.278	0.925	49.7	0.052
	\mathcal{L}_1 -DISTS	0.826	0.557	0.852	16.91	0.844	0.291	0.927	133.9	0.043
	CONTENT	0.848	0.565	0.858	17.05	0.849	0.298	0.930	165.8	0.075
	\mathcal{L}_1 -CONTENT	0.824	0.562	0.855	17.01	0.846	0.296	0.929	176.8	0.086
Maps	\mathcal{L}_1	0.687	0.515	0.728	20.24	0.623	0.083	0.885	169.1	0.084
	LPIPS	0.686	0.517	0.731	20.26	0.628	0.086	0.885	163.1	0.079
	\mathcal{L}_1 -LPIPS	0.693	0.518	0.728	20.28	0.631	0.085	0.888	162.8	0.081
	DISTS	0.674	0.487	0.710	20.03	0.607	0.078	0.880	144.4	0.067
	\mathcal{L}_1 -DISTS	0.682	0.509	0.724	20.27	0.621	0.083	0.883	147.3	0.068
	CONTENT	0.685	0.497	0.701	19.75	0.619	0.082	0.878	174.1	0.093
	\mathcal{L}_1 -CONTENT	0.689	0.516	0.725	20.16	0.622	0.082	0.885	171.3	0.093
Papsmear	\mathcal{L}_1	0.832	0.581	0.917	20.85	0.874	0.121	0.886	72.36	0.036
	LPIPS	0.841	0.576	0.922	20.82	0.877	0.144	0.886	39.41	0.002
	\mathcal{L}_1 -LPIPS	0.840	0.592	0.924	20.85	0.879	0.148	0.887	47.91	0.013
	DISTS	0.832	0.589	0.913	20.83	0.869	0.110	0.886	45.80	0.009
	\mathcal{L}_1 -DISTS	0.831	0.601	0.912	20.85	0.868	0.110	0.888	46.57	0.003
	CONTENT	0.824	0.568	0.910	20.90	0.863	0.107	0.891	75.94	0.041
	\mathcal{L}_1 -CONTENT	0.818	0.581	0.911	20.88	0.858	0.102	0.886	82.47	0.049

The fourth dataset analyzed in the article is Papsmear. That is a medical dataset. It differs from the Cityscapes and Maps datasets. The objects (nucleus, cytoplasm, etc.) in the images in the Papsmear dataset are partitioned interconnectedly. Figure 5 presents the visual results of the Papsmear dataset to understand the outcome of the other losses of the Pix2Pix. Looking at the up-to-date similarity metrics for the Papsmear dataset in Table 1, it is observed that the LPIPS function gives better results than other methods. Looking at the conventional similarity metrics FSIM, MS-SSIM, SSIM, and VIFp in Table 1, the \mathcal{L}_1 _LPIPS function shows; the \mathcal{L}_1 _DISTS function in HaarPSI; It has been observed that the CONTENT function gives good results in PSNR and VSI.

As a result, when the translation results in the datasets are evaluated in general, it is seen that DISTS and LPIPS functions provide adequate accuracy compared to the other metrics. It has been observed that conventional similarity measures cannot provide sufficient accuracy. In contrast, up-to-date similarity measures give better results in the DISTS function for the Cityscapes and Maps datasets and the LPIPS function for the Denim2Mustache and Papsmear datasets. Consistent similarity results from DISTS and LPIPS functions are similar to the study results (Ding et al., 2021). Thus, it has been seen that DISTS and LPIPS functions can be used for loss measurement for GANs architectures. Up-to-date similarity metric have been studied in detail as they give more accurate results. In the Cityscapes dataset, the DISTS function showed success with 67.77 FID and 0.036 KID values. In the Maps dataset, the DISTS function achieved an FID of 144.4 and a KID of 0.067. For the Denim2Mustache dataset, the LPIPS function demonstrated success with an FID of 130.7 and a KID of 0.042. Notably, in the Papsmear dataset, the FID and KID results were 39.47 and 0.002, respectively, and it was observed that the LPIPS function outperformed the other functions.

In this study, the original \mathcal{L}_1 loss function has been added to some Convolutional Neural Network methods in addition to the standard Generative Adversarial Network (GAN) architectures, specifically the Pix2Pix method. The purpose of keeping the GAN method constant is to observe the impact of the \mathcal{L}_1 loss function on Convolutional Neural Networks. When the \mathcal{L}_1 function is added, the model is tested on the Pix2Pix method, which is a fixed GAN architecture. When the proposed methods, namely \mathcal{L}_1 _DISTS and \mathcal{L}_1 _LPIPS, are examined metrically, they are observed to achieve better performance. These methods are advantageous as they reach the result faster and more accurately. In summary, the addition of the \mathcal{L}_1 loss function leads to higher performance. It has been observed that in the future, the \mathcal{L}_1 loss function may be used in addition to other methods. In some cases, such as variations in the dataset, visual observations indicate that there is variability in the results, and in certain situations, the \mathcal{L}_1 loss function is visually observed to be less successful.

4. Conclusion

The aim of this study is to evaluate the performance of loss function on the Pix2Pix architecture for GANs. CNN-based loss functions (CONTENT, DISTS, and LPIPS) were used instead of Pix2Pix's original \mathcal{L}_1 loss. Four different datasets were used to examine the effect of the loss function. The effects of adding CNN-based structures to the contentious loss term and regularization terms in the loss function are analyzed. As a result of the training and testing process, translation accuracies were transferred to tables with conventional and up-to-date metrics. When the experimental results were examined, it was seen that the LPIPS and DISTS method had the best synthesis performance according to the up-to-date similarity metrics (FID and KID). It seems that conventional similarity measures do not give consistent results in the translating accuracy. It is seen that the DISTS function in datasets with high complexity Cityscapes, Maps, and the LPIPS function with less complexity. Denim2Mustache and Papsmear give better results compared to other methods. As a result, it can be said that using DISTS and LPIPS functions in image-to-image translation architectures positively effects the translating accuracy.



Figure 2. Synthesis results of the Cityscapes dataset. Rows: Real image, Ground-truth, \mathcal{L}_1 , LPIPS, \mathcal{L}_1 -LPIPS, DIST, \mathcal{L}_1 -DIST, CONTENT, \mathcal{L}_1 -CONTENT

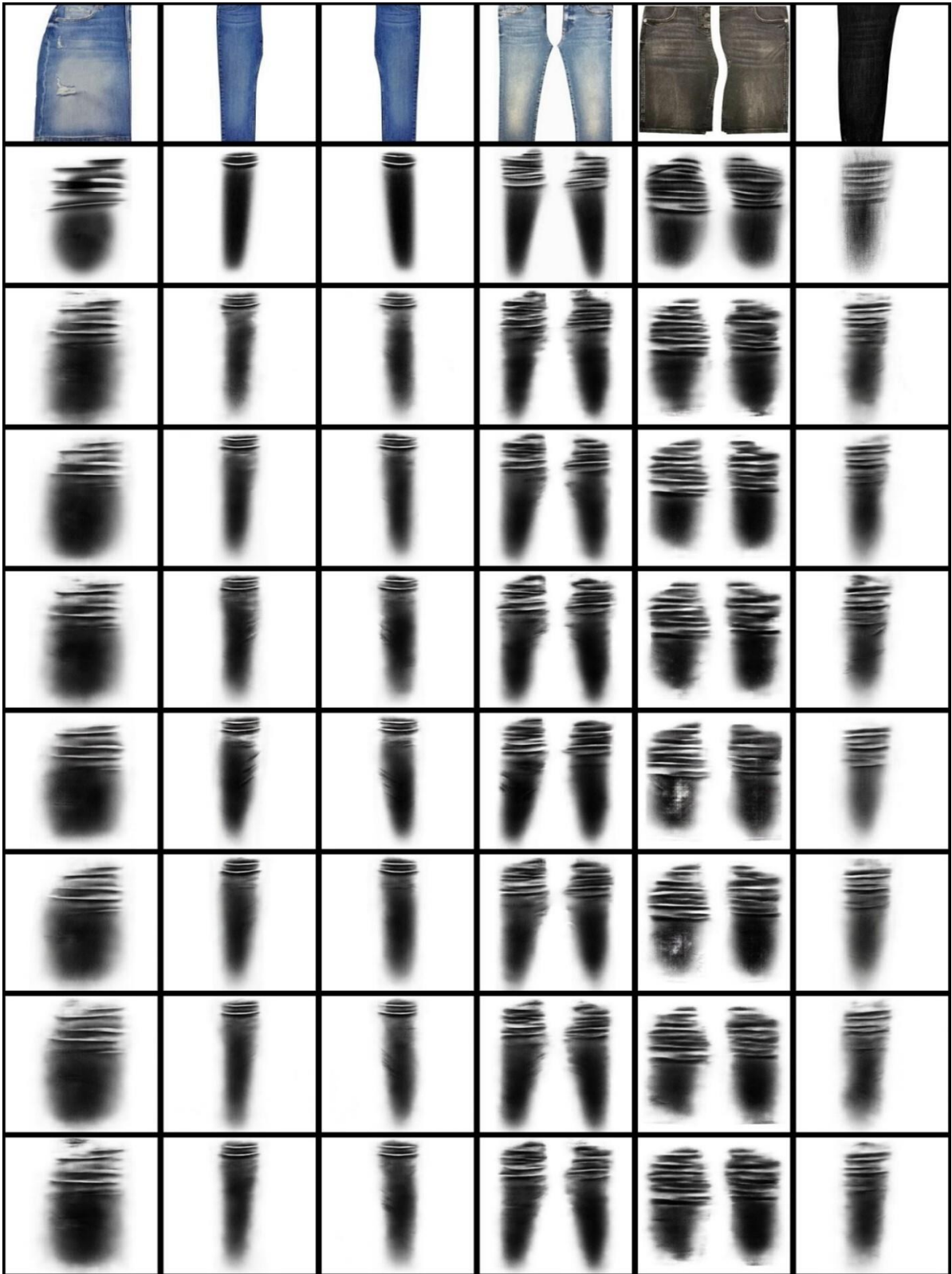


Figure 3. Synthesis results of the Denim2Mustache dataset. Rows: Real image, Ground-truth, \mathcal{L}_1 , LPIPS, \mathcal{L}_1 -LPIPS, DISTS, \mathcal{L}_1 -DISTS, CONTENT, \mathcal{L}_1 -CONTENT

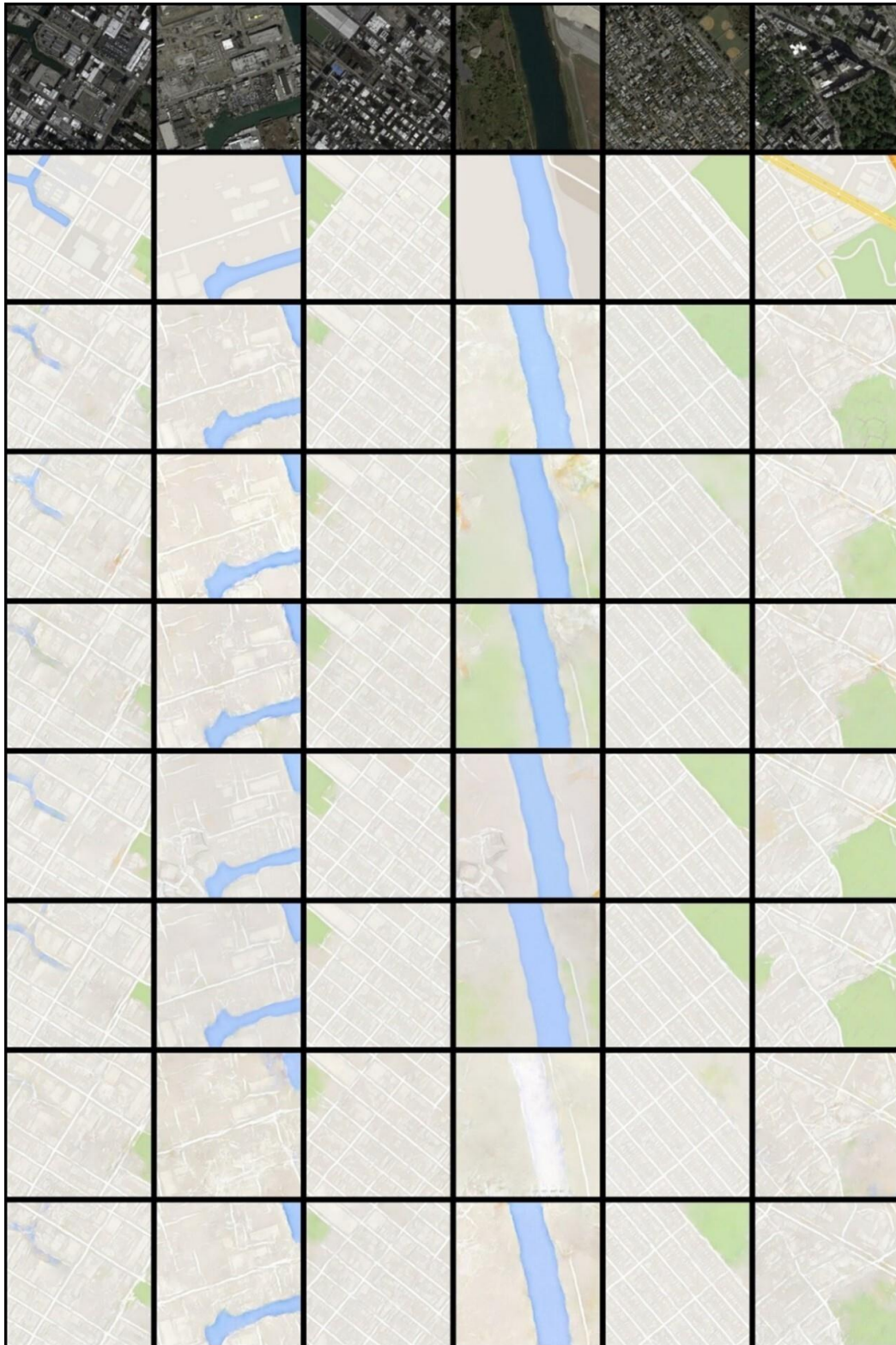


Figure 4. Synthesis results of the Denim2Mustache dataset. Rows: Real image, Ground-truth, \mathcal{L}_1 , LPIPS, \mathcal{L}_1 _LPIPS, DIST, \mathcal{L}_1 _DIST, CONTENT, \mathcal{L}_1 _CONTENT

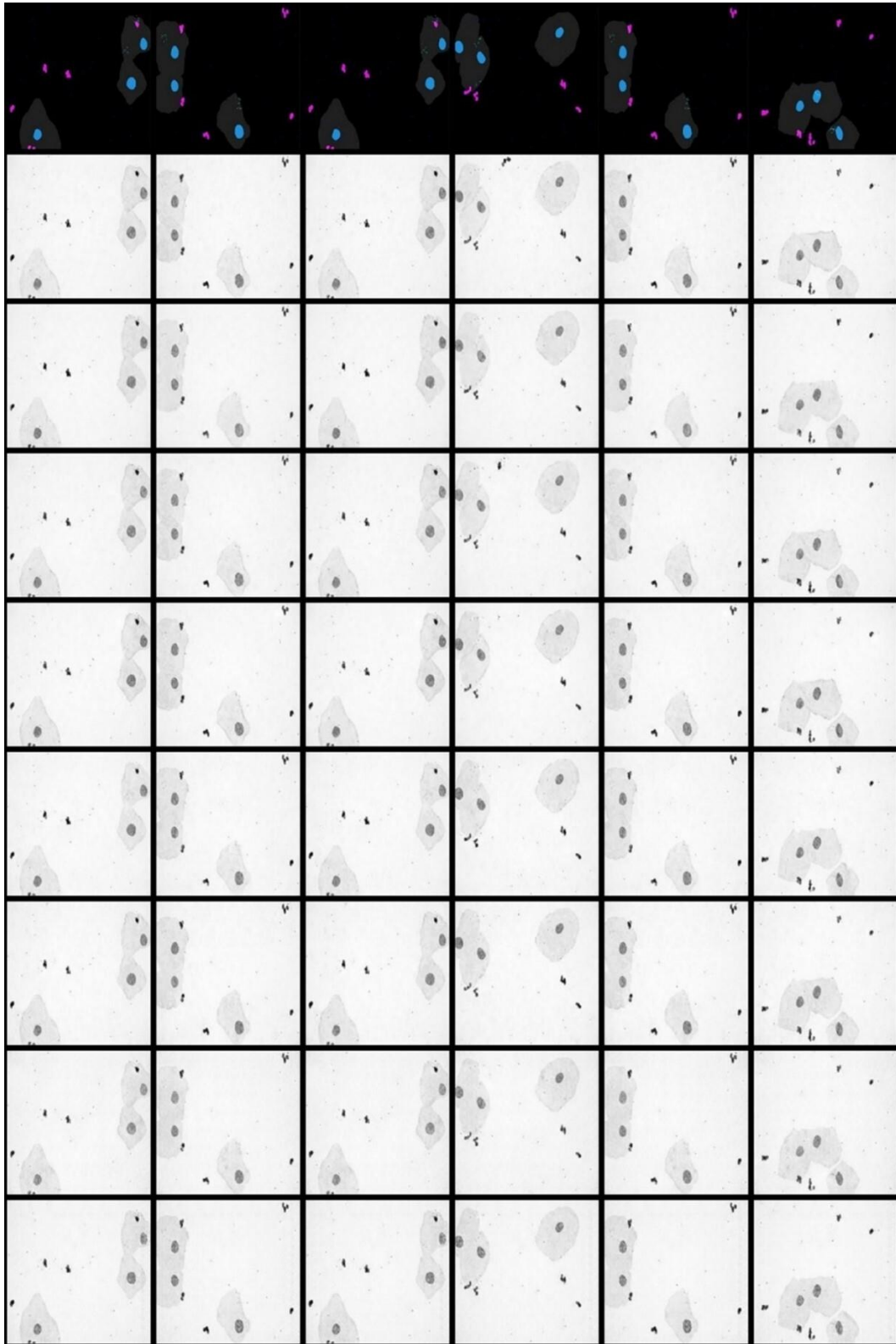


Figure 5. Synthesis results of the Denim2Mustache dataset. Rows: Real image, Ground-truth, \mathcal{L}_1 , LPIPS, \mathcal{L}_1 -LPIPS, DIST, \mathcal{L}_1 -DIST, CONTENT, \mathcal{L}_1 -CONTENT

References

- Zhu, X. X., Tuia, D., Mou, L., Xia, G. S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience and remote sensing magazine*, 5(4), 8-36.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).
- Koushik, J. (2016). Understanding convolutional neural networks. *arXiv preprint arXiv:1605.09081*.
- Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, 187, 27-48.
- Van Den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016, June). Pixel recurrent neural networks. In *International conference on machine learning* (pp. 1747-1756). PMLR.
- Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., & Graves, A. (2016). Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.
- Salimans, T., Karpathy, A., Chen, X., & Kingma, D. P. (2017). Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*.
- Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017). Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5659-5667).
- Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30.
- Liu, M. Y., & Tuzel, O. (2016). Coupled generative adversarial networks. *Advances in neural information processing systems*, 29.
- Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Wang, T. C., Liu, M. Y., Zhu, J. Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8798-8807).
- Liu, Y., De Nadai, M., Yao, J., Sebe, N., Lepri, B., & Alameda-Pineda, X. (2020). Gmm-unit: Unsupervised multi-domain and multi-modal image-to-image translation via attribute gaussian mixture modeling. *arXiv preprint arXiv:2003.06788.1*
- Royer, A., Bousmalis, K., Gouws, S., Bertsch, F., Mosseri, I., Cole, F., & Murphy, K. (2020). Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding* (pp. 33-49). Cham: Springer International Publishing.
- Royer, A., Bousmalis, K., Gouws, S., Bertsch, F., Mosseri, I., Cole, F., & Murphy, K. (2020). Xgan: Unsupervised image-to-image translation for many-to-many mappings. In *Domain Adaptation for Visual Understanding* (pp. 33-49). Cham: Springer International Publishing.
- Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321, 321-331.
- Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1125-1134).
- Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).
- Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 586-595).

- Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5), 2567-2581.
- Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.
- Zhang, L., Zhang, L., Mou, X., & Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8), 2378-2386.
- Reisenhofer, R., Bosse, S., Kutyniok, G., & Wiegand, T. (2018). A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61, 33-43.
- Fardo, F. A., Conforto, V. H., de Oliveira, F. C., & Rodrigues, P. S. (2016). A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms. *arXiv preprint arXiv:1605.07116*.
- Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003, November). Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003 (Vol. 2, pp. 1398-1402)*. Ieee.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4), 600-612.
- Sheikh, H. R., & Bovik, A. C. (2006). Image information and visual quality. *IEEE Transactions on image processing*, 15(2), 430-444.
- Zhang, L., Shen, Y., & Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image processing*, 23(10), 4270-4281.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Bińkowski, M., Sutherland, D. J., Arbel, M., & Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.
- Choi, Y., Uh, Y., Yoo, J., & Ha, J. W. (2020). Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 8188-8197)*.
- Ding, K., Liu, Y., Zou, X., Wang, S., & Ma, K. (2021, October). Locally adaptive structure and texture similarity for image quality assessment. In *Proceedings of the 29th ACM International Conference on Multimedia (pp. 2483-2491)*.
- Sim, K., Yang, J., Lu, W., & Gao, X. (2020). MaD-DLS: mean and deviation of deep and local similarity for image quality assessment. *IEEE Transactions on Multimedia*, 23, 4037-4048.
- Borasinski, S., Yavuz, E., & Béhuret, S. (2022). Paired Image-to-Image Translation Quality Assessment Using Multi-Method Fusion. *arXiv preprint arXiv:2205.04186*.
- Peng, X., Peng, S., Hu, Q., Peng, J., Wang, J., Liu, X., & Fan, J. (2022). Contour-enhanced CycleGAN framework for style transfer from scenery photos to Chinese landscape paintings. *Neural Computing and Applications*, 34(20), 18075-18096.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Suzuki, A., Akutsu, H., Naruko, T., Tsubota, K., & Aizawa, K. (2021). Learned Image Compression with Super-Resolution Residual Modules and DISTS Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1906-1910)*.
- Chuan, P. M., Son, L. H., Ali, M., Khang, T. D., Huong, L. T., & Dey, N. (2018). Link prediction in co-authorship networks based on hybrid content similarity metric. *Applied Intelligence*, 48, 2470-2486.
- Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18 (pp. 234-241)*. Springer International Publishing.

- Li, C., & Wand, M. (2016). Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14* (pp. 702-716). Springer International Publishing.
- Ding, K., Ma, K., Wang, S., & Simoncelli, E. P. (2021). Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, 129, 1258-1281.
- Mihelich, M., Dognin, C., Shu, Y., & Blot, M. (2020, June). A characterization of mean squared error for estimator with bagging. In *International Conference on Artificial Intelligence and Statistics* (pp. 288-297). PMLR.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3213-3223).
- Şahin, E., & Talu, M. F. (2021). Bıyık Deseni Üretiminde Çekişmeli Üretici Ağların Performans Karşılaştırması. *Bitlis Eren Üniversitesi Fen Bilimleri Dergisi*, 10(4), 1575-1589.
- Altun, S., & Talu, M. F. (2022). A new approach for Pap-Smear image generation with generative adversarial networks. *Journal of the Faculty of Engineering and Architecture of Gazi University*, 37(3), 1401-1410.