Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-14090-6

Popular image generation based on popularity measures by generative adversarial networks

Narges Mohammadi Nezhad[1] · **Seyedeh Leili Mirtaheri**[2] · **Reza Shahbazian**[3]

Received: 20 August 2021 / Revised: 15 March 2022 / Accepted: 21 October 2022

Abstract
We study image-to-image translation and synthetic image generation. There is still no developed model to create popular synthetic images based on the user's opinion in the fashion industry. This paper uses a combination of generative adversarial networks (GAN), deep learning, and user's opinions to create popular images. Our proposed model consists of two modules; one is a popularity module that estimates the intrinsic popularity of images without considering the effects of non-visual factors. The second one is a translation module that converts unpopular images into popular ones. Our model also performs multi-dimensional translation and multi-domain translation. We use the ResNet50 neural network as the default deep neural network in which the last layer is replaced with a fully connected layer. We use a new dataset collected from Instagram to train our network. We evaluate the performance of the proposed method by FID, LPIPS scores, and popularity index in different scenarios. The results show that our proposed method shows at least 60% and 25% improvement in terms of FID and LPIPS in color-to-color image translation. These improvements confirm the proposed method's generated images' quality and diversity. The evaluations on the popularity score also confirms that the content-based translation is more effective than style-based translation in terms of popularity.

Keywords Image popularity · Generative adversarial network (GAN) · Artificial intelligence · Deep learning · Image translation · Synthetic

1 Introduction

Artificial intelligence, computer vision, and deep learning methods have applications in various fields of daily life from automatic care of elders [2, 37] to chair design [34], language

✉ Seyedeh Leili Mirtaheri
Mirtaheri@khu.ac.ir

[1] Department of Computer Science, Faculty of Mathematics and Computer Science, Kharazmi University, Tehran, Iran

[2] Electrical, Computer Engineering, Faculty of Engineering, Kharazmi University, Tehran, Iran

[3] Electrical Engineering Research Group, Faculty of Technology and Engineering Research Center, Standard Research Institute, Alborz, Iran

generation [31], IoT wearable devices [1, 3], voice conversion [21] and image synthesis [35]. One of the interesting topics in computer vision is digital image generation [16]. The image-to-image translation is a class of vision problems in which the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. The image-to-image translation is used in various contexts and applications [6, 7, 13, 14, 23, 29, 38, 46]. Automatic image-to-image translation is the problem of translating one possible representation of an image or scene into another, given sufficient training data [36].

Artificial intelligence, computer vision, and deep learning methods have applications in various fields of daily life from automatic care of elders [2, 37] to chair design [34], language generation [31], IoT wearable devices [1, 3], voice conversion [21] and image synthesis [35]. One of the interesting topics in computer vision is digital image generation [16]. The image-to-image translation is a class of vision problems in which the goal is to learn the mapping between an input image and an output image using a training set of aligned image pairs. The image-to-image translation is used in various contexts and applications [6, 7, 13, 14, 23, 29, 38, 46]. Automatic image-to-image translation is the problem of translating one possible representation of an image or scene into another, given sufficient training data [36].

New tools and models exist for automatic image-to-image translation, including the Generative Adversarial Networks (GAN) and conditional GAN [36]. Using GAN to generate images is well studied in the literature [8, 11, 22–24, 28, 30, 32, 40–42, 47, 49, 51]. image-to-image translation aims at learning the relationship between samples from two image domains. The relationship between two domain images can be one-to-one, one-to-many, or many-to-many. In a one-to-many mapping image translation problem, an input sample from one domain can correspond to multiple samples in the other domain. The existing models for image-to-image translation, focus on learning a one-to-many mapping. These existing models cannot generate images from both aspects of multi-dimensional translation at the same time. These methods also do not consider the user's taste; that is the main idea behind our proposed model. We also propose popularity metrics on GAN-based generated images for further evaluation.

The image-to-image technology can be used in industrial topics, including fashion and clothing, while it is neglected in the literature. In this paper, we use image-to-image translation for the fashion industry. We collect and use a dataset of clothes along with the existing metadata such as likes to extract the popularity metrics of clothes. Using the extracted metadata, we apply the user's taste to create popular synthetic images. In summary, the novelties and contributions of this paper are listed as follows:

- We study and review different models for image translation and generation by using GANs and their potential application in the fashion industry. We introduce the gap in the literature and propose to add the user's opinion as a key factor in synthetic image generation and improve the synthetic image desires accordingly.
- We create a dataset by collecting images from Instagram. Our dataset includes clothes pictures along with the number of likes and existing metadata.
- We use GAN and propose a model to generate popular images for the fashion industry. Our proposed model consists of two modules; one is a popularity module that estimates the intrinsic popularity of images without considering the effects of non-visual factors. The second one is a translation module that converts unpopular images into popular ones. Unlike existing methods that use different discriminators for different domains, our proposed method, in its second module, adopts DMIT [48] and uses a unified conditional discriminator for all domains. Our model also performs multi-dimensional translation and multi-domain translation.

– We evaluate the performance of the proposed method by considering FID and LPIPS scores to explore the quality and diversity of the generated images. We also extract the popularity metrics and evaluate the performance of content and style-based image-to-image generation on the popularity score. By applying the metrics of popular images, we train our model to generate synthetic images with the specifications of popular images. Our model considers the users' opinions.

The rest of this paper is organized as follows: in Section 2, we review the related works. Section 3 presents our proposed method for improving and evaluating Image Generation and image-to-image translation. Evaluation metrics and experimental results are presented in Section 4, and Finally, Section 5 concludes the paper.

2 Related work

2.1 Generative adversarial network

GAN has achieved great success in the field of image generation since its presentation in 2014 [15]. GAN consists of two neural networks, named generator and discriminator, and solves the MinMax problem. Based on the MinMax problem, the Generator network is trained to deceive the discriminator, while the discriminator learns to distinguish the generator's fake images from the real ones. Adversarial networks do not need Markov chains and use Back-propagation to obtain gradients, which is advantageous.

GANs can generate new images instantly based on the learned distribution, while the recurrent neural networks work pixel-to-pixel. One of the most critical and specific advantages of a GAN is that the performance of discriminator D is very similar to a trainable loss and is compatible with different tasks and datasets. This adaptive loss enables GANs to solve many challenging image generation and categorization problems.

Authors in [39] use GAN to capture the common and salient objects in a group of relevant images, called co-saliency detection. To improve the quality of the generated images, various GAN network models have been proposed, including the DCGAN [4] and PGAN [25]. Authors in [10] propose an effective defense framework specified for remote sensing image scene classification, named perturbation-seeking generative adversarial networks (PSGANs). Their proposed framework is designed to train the classifier by introducing the examples generated during the reconstruction process of images. These generated examples can be random kinds of unknown attacks during training and thus are utilized to eliminate the blind spots of a classifier. They model the distributions of the perturbations added in adversarial examples. Although this method is also considered in GAN-based image generation, it is a bit different by the common understanding of image-to-image translation.

2.2 Image-to-image translation

Various models of GAN-based image-to-image translation models have been proposed. The first model proposed for image-to-image translation is Pix2Pix [23]. This framework uses paired data to translate image to image by using a conditional GAN. The BicycleGAN [24] trains by a multi-modal mapping of paired data for supervised image-to-image translation. The main idea of BicycleGAN is based on objective consistency and a multi-modal that combines both mappings between latent space and target space, and despite its complex learning process, it can generate more diverse and realistic results compared with Pix2Pix

presented in [23]. Although BicycleGAN can produce multi-modal results, the quality and variety of images generated are still far from desirable.

In translating images, the critical challenge is to learn the joint distribution of images in different domains. The UNIT [20] model is introduced to solve this problem [32]. This model is based on the theory of shared latent space between domains of two images. Training of this network is unsupervised and, therefore, requires unpaired data. This model can generate high-quality images; however, its training process is unstable and difficult. Besides, this model is uni-modal due to the Gaussian latent space theory [28, 40].

The MUNIT [22] model was introduced for image-to-image translation without supervision. In this network, the latent space of images is decomposed into a content space, which is a fixed domain, and a style space that includes specific features of the domain and generates multi-dimensional outputs. In methods such as BicycleGAN and pix2pix, pairs of images are needed to train the network in a supervised manner [23, 24]. However, the MUNIT model training is unattended, and paired images are needed. One of the limitations and problems of MUNIT is that this network is limited to only two domains at a time and does not perform multi-domain styles. Besides, this network does not have an integrated structure. As a result, the generated images by MUNIT do not seem real enough.

StarGan and StarGan v2 [11] perform the multi-domain translation and use noise vectors to generate more diverse results by the generator. The StarGan model uses only an integrated structure for learning two domains mapping. The generated images by StarGan do not look real enough, and the model suffers from the mode collapse problem [42]. In StarGan v2, the quality of the images generated has improved due to the use of noise distribution. However, the results are still not very diverse. DRIT [30] and SingleGAN [49] generate real results, although the images are not clear enough.

Previous studies in the image-to-image translation focus on learning a one-to-many mapping from both aspects of multi-dimensional translation. However, they cover only one of the two modes at each time and cannot generate images from both elements at the same time [11, 22–24, 28, 30, 32, 40, 42, 49].

The image-to-image translation can consider users' tastes and interests, although it is not addressed in the literature. In the next section, we propose a model that incorporates the users' preferences in generating and translating images in terms of content and style.

3 Proposed method

Assume that we aim to design (generate) a T-Shirt with popular sufficient shape and popular texture variation while maintaining high resolution. We introduce the image-to-image (I2I) translation module and the popularity module to solve the above-mentioned issues, respectively. The pipeline of our proposed method is illustrated in Fig. 1.

At first, we use the popularity module on the dataset. The dataset used in this paper is collected from Instagram and explained in Section 4.1. We consider intrinsic evaluation metric [12] and divide the dataset images into two categories, popular and non-popular. Further, these two categories become two inputs for the second module, image-to-image Translation (Generation). We train the second module network with popular and non-popular categories. The content and style network learn popular and non-popular images and can generate images, accordingly. In the second module, different and separate loss functions are defined for content and style, and the generator network learns to use these loss functions to generate images with popular content and style. Finally, the generated images by using the second

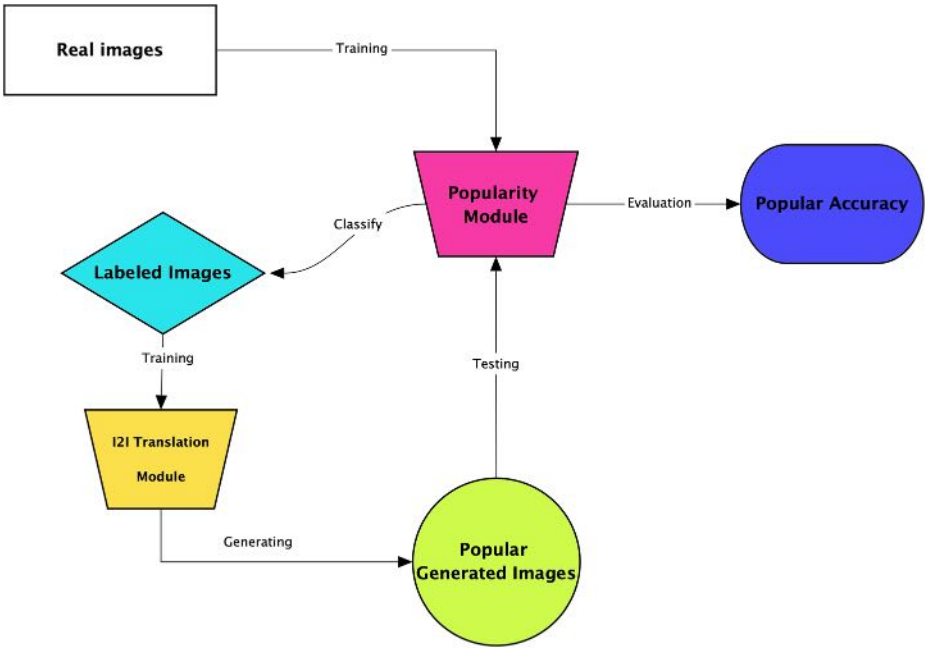


Fig. 1 The pipeline of the proposed model includes two modules: Popularity and I2I Translation. The first Popularity Module is trained on our collected dataset, and then I2I Translation Module is applied to the input

module are given to the first module for evaluating the popularity of the generated images with the intrinsic evaluation metric.

3.1 Popularity module

This module estimates the intrinsic popularity of images without considering the effects of non-visual factors on the popularity of the image. The results show that this model predicts the intrinsic popularity of images more accurately compared with the previous methods [19, 26]. As illustrated in Fig. 2, to calculate the popularity of images, we first generate an image database that includes data popularity discriminable image pairs (PDIPs) and then apply the deep neural network (DNN) computational model to the dataset.

We use the ResNet50 neural network [18] as the default deep neural network by modifying its architecture to evaluate the collected dataset. In the ResNet50 model, the last layer is replaced with a fully connected layer. The input images' size is set to 256x256, and we use the Adam function [27] for optimization with a batch size of 64.

$$Q_A = f(A) \tag{1}$$

$$Q_X = f^*(X) \tag{2}$$

$$Q_{AB} = f(A) - f(B) \tag{3}$$

$$P_{AB} = \frac{\exp(Q_{AB})}{1 + \exp(Q_{AB})} \tag{4}$$

The popularity score ranges from -2 to 3. The higher the value, the more popular the image is.

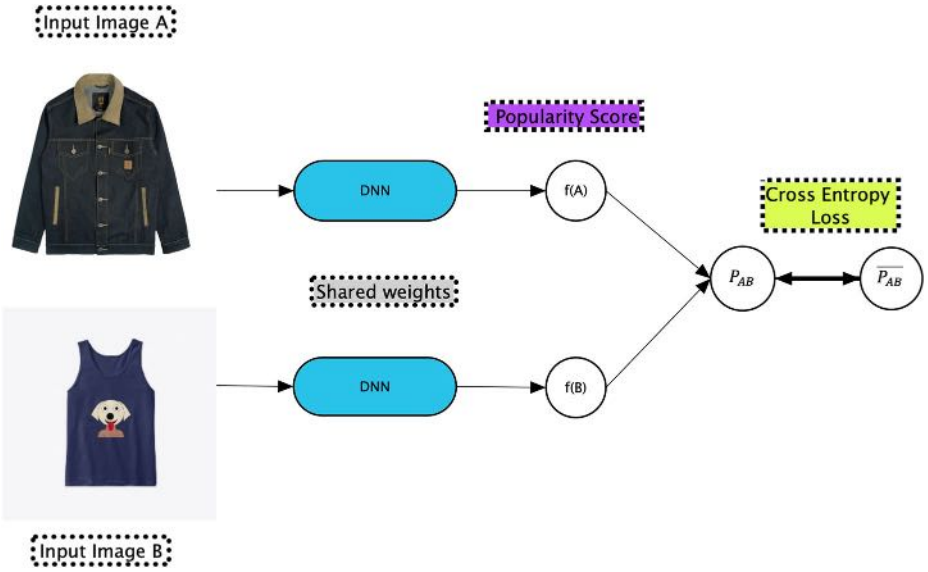


Fig. 2 Structure of the popularity module (Intrinsic). The computational model is based on a deep neural network (DNN) for estimating the popularity score of two images

3.2 Image-to-image translation (generation) module

The image-to-image translation module is based on DMIT [48] and aims to learn the distribution of clothing datasets collected from Instagram images and generate synthetic images accordingly. DMIT [48] is developed based on the theory of shared latent space [33], which is divided into two separate parts: modeling a multi-model distribution and obtaining a cross-domain translation. In this model, images are separated into two latent space representations, including a latent space C and a Style space S, and use an encoder and decoder to learn disentangled representations. This module consists of the following parts:

1. The style encoder E_S Extracts the style from the input image.
2. Content encoder E_C is a fully convolutional network that encodes the input image into a C feature space representation and represents the input image to its content.
3. D module is defined as a set of domain labels and treats D as another disentangled representation of the images.

We use the I2I module as an unsupervised integrated network that includes; a unified Generator and a unified conditional discriminator. The proposed module performs one-to-many mapping in image-to-image translation from two aspects; multi-dimensional translation and multi-domain translation. The conditional adversarial training forces the model to generate realistic images with a wide variety. We also define a mapping function between x and D called a domain label encoder with a dictionary structure that extracts the input images' domain labels. Using Generator G, we can perform simultaneous modeling for multi-domain and multi-dimensional translations. As a result, with any desired style s member of S and domain label d member of D, we can translate an input image x member of X to the corresponding target x_t member of X.

$$X_t = G(E_C(X_i), S, d) \tag{5}$$

Unlike previous methods that use different discriminators for different domains, this module uses a unified conditional discriminator for different domains.

3.3 Training method

The learning process is divided into the translation path and the disentanglement path.

Disentanglement path This path is like an encoder-decoder architecture that uses conditional adversarial training on the latent space. In this path, encoders encode the image into disentanglement representations that can be returned to the conditional generator's input image. The disentanglement path uses CVAE [43] as the base structure to separate the latent spaces of the images. To align style representations across visual domains and limit style information, we bring all domains' style distribution to an initial distribution, as close as possible. The schematic of the disentanglement path is depicted in Fig. 3.

$$\begin{aligned}\mathcal{L}_{cVAE} &= \lambda_{KL} E_{x_i \sim X}\left[KL(E_s(x_i) \parallel q(s)\right] \\ &+\lambda_{rec} E_{x_i \sim X}\left[\|G(E_c(x_i), E_s(x_i), (E_d(x_i)) - x_i\|_1\right]\end{aligned} \tag{6}$$

To enable random sampling at test time, we select the initial distribution $q(s)$ as the standard Gaussian distribution $N(0, 1)$. For content representation, we suggest performing conditional adversarial training in the content space to address the issue of changing the distribution of content on domains. This process encourages deletion of domain d information in content c.

$$\begin{aligned}\mathcal{L}^C_{GAN} &= E_{x_i \sim X}\left[\log(D_c(E_c(x_i), E_d(x_i)))\right] \\ &+E_{d \sim (D - \{E_d(x_i)\})}\left[\log(1 - D_c(E_c(x_i), d))\right]\end{aligned} \tag{7}$$

The total loss of the disentanglement path is calculated as follows:

$$\mathcal{L}_{D-Path} = \mathcal{L}_{cVAE} + \mathcal{L}^C_{GAN} \tag{8}$$

Without a D-path, the photos generated are blurry and unrealistic, and their diversity is very low.

Translation path Path D persuades the model to learn C content and S style with an initial distribution. The schematic of the Translation path is presented in Fig. 4. However, two issues still remain;

1. Due to the limited amount of training data and the optimization of KL error, the Generator, G may only sample a subset of S and generate images with specific domain tags during the training phase; that may result poorly in the generation when sampling S in the initial N and D distributions do not match the test image [45].
2. The above training process lacks efficient motives to use styles, which leads to the low diversity of generated images.

To solve the above-mentioned two problems and persuade the generator to obtain a full distribution of the outputs, we first randomly sample the domain and style tags in the prior

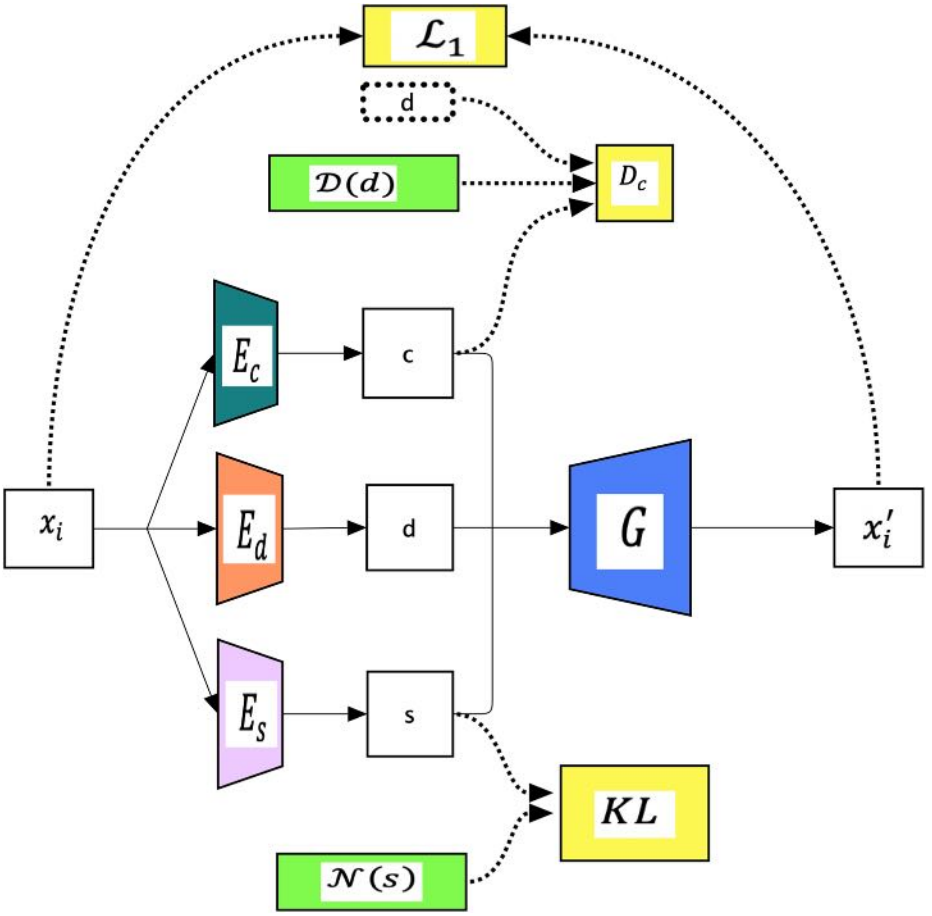


Fig. 3 The schematic of the Disentanglement Path. This path is like an encoder-decoder architecture that uses conditional adversarial training. In this path, disentanglement representations can be returned to the conditional generator's input image

distribution to cover the entire sampling space during training. Then latent regression [9, 52] is used to force the generator to use the style vector. Regression can be performed on the content of C to separate the style S from C. The latent regression is re-written as follows:

$$\begin{aligned}\mathcal{L}_{reg} = & E_{\substack{c\sim C\\ s\sim N\\ d\sim D}}\left[\|E_s(G(c,s,d))-s\|_1\right] \\ & +E_{\substack{c\sim C\\ s\sim N\\ d\sim D}}\left[\|E_c(G(c,s,d))-c\|_1\right] \end{aligned} \tag{9}$$

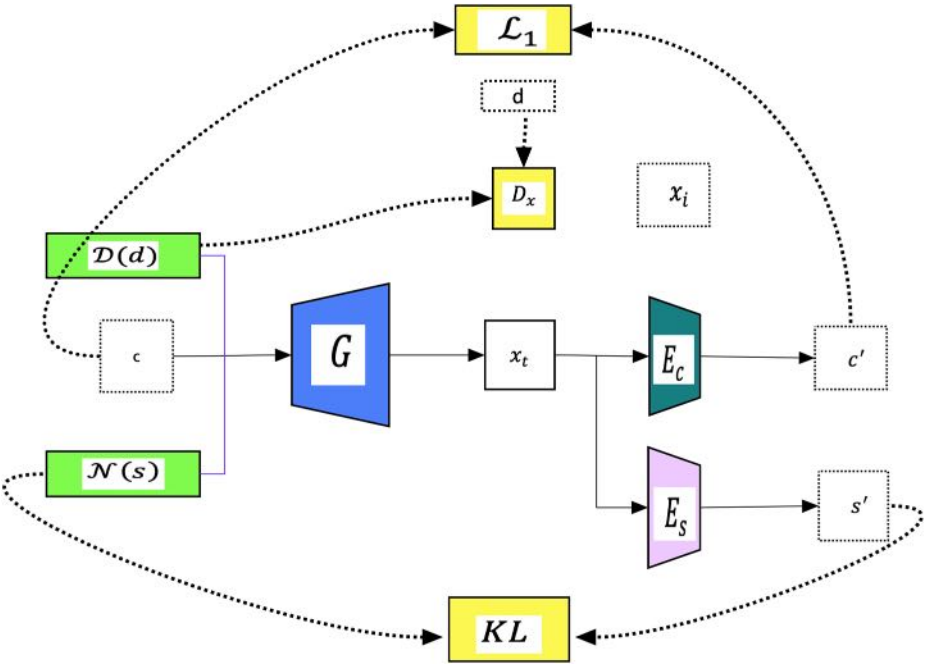


Fig. 4 The schematic of the Translation path. In this path, D persuades the model to learn C content and S style with an initial distribution

To match the distribution of the generated images with real data between sampling domain labels and styles, we use conditional adversarial training in pixel space:

$$\begin{aligned}\mathcal{L}^{x}_{GAN} = & E_{x_i \sim X}\left[\log(D_x(x_i, E_d(x_i)))\right] \\ & + E_{d\sim(D-\{E_d(x_i)\})}\left[\frac{1}{2}\log(1 - D_x(x_i, d))\right] \\ & + E_{s\sim\mathcal{N}}\left[\frac{1}{2}\log(1 - D_x(G(E_c(x_i), s, d), d)\right]\end{aligned} \tag{10}$$

To encourage the generator to generate images that fit the given domain label, we separate the actual image from the unsuitable target domain. The final target of the translation is as follows:

$$\mathcal{L}_{T-Path} = \lambda_{reg}\mathcal{L}_{reg} + \mathcal{L}^{x}_{GAN} \tag{11}$$

By combining both training paths, the final objective function of the model is as follows:

$$\min_{G,E_c,E_s}\max_{D_c,D_x} \mathcal{L}_{D-Path} + \mathcal{L}_{T-Path} \tag{12}$$

4 Experiment results

4.1 Dataset

To do this research, we needed a diverse collection of data that included high-quality images alongside the number of likes. There exist some methods for explicit and implicit fuzzy like

in social media that can be used to realize their like level and show relevant advertisements [17]. After conducting various researches, we realized that we do not have the required dataset with these specifications, and therefore, we collected the desired dataset from Instagram. We reviewed the various pages of Instagram and to find images of clothes with a white background that have no other object to make sure that the images are popular only because of the clothes, not the other objects and nor scenery of the image. In translation, we need single-object images containing only one garment that does not include humans or other elements.

Reviewing many pages with a high number of followers and audiences and inspired by the method introduced by Ding et.al in 2019 [12], we were able to find the first suitable page called "Konglomerads". We continued the data collection benefiting from Instagram's recommender system on similar pages. Finally, we selected twenty-one Instagram profiles containing the images with required features. Statistics and specifications of these 21 profiles, along with the number of followers and the number of posts, are summarized in Table 1.

Our dataset consists of images along with the number of likes, download time, post address, user ID, content type, image post time, caption including the hashtag, and the number of comments for each photo. In this paper, the collected data are only the images of the posts, and the collection of other data types such as video is omitted. In total, we collected around 130,000 images.

The database needs to meet the following conditions:

1. The intrinsic popularity of image A in the database is greater than the intrinsic popularity of image B, which sets a threshold level of T.
2. Each image has been published for at least one month because researches show that the number of likes does not increase considerably after one month.
3. The image's caption must have a maximum of 6 words, a hashtag, and without @sign. This prevents false popularity because of images being tagged by different people in the caption or comments.

First, we pre-process the collected images. In this step, we remove all duplicates and the images, including objects beside the clothes. We resized all the images to 256x256. After pre-processing, our dataset includes 30,000 images. Out of 130,000 collected images, 30,000 met the conditions above. We label the images as popular and non-popular, according to the defined parameters in [12].

We used 27,000 images for the training data and 3,000 for the test dataset. These images are collected with appropriate quality and diversity.

To evaluate the proposed model accurately, we first convert the images into black-and-white colors that contain only clothing design lines and edge images of the dataset. The size of the converted edge images is set to 256x256. We also design all the clothes' lines so that the converted images are authentic and close to the original images.

4.2 Implementations

The overall architecture of the proposed method is shown in Fig. 1. For image translation, we adopt DMIT [48] architecture that consists of generator G with several residual blocks, followed by several de-convolutional layers. Each convolution layer in residual blocks is equipped with CBIN [48] for information injection, discriminator, content encoder, and style encoder. For popularity, we adopt the architecture of Intrinsic (CNN Model) [12] with

Table 1 The details of the collected dataset from Instagram, including the 21 profiles along with the number of followers and the number of posts

Index	Profile Name	Number of Followers	Number of Posts
1	konglomerads	32,400	3,891
2	screamous_store	426,000	8,098
3	animous.store	4,980	215
4	efg.catalog	285,000	446
5	itsmailmo.catalog	38,000	265
6	blankwear_catalog	45,600	4,893
7	vearst_storage	88,700	44
8	evilarmyshop	148,000	411
9	thxnsmncatalog	653,000	260
10	maternaldisaster	592,000	4,957
11	coloreswear	1,800	603
12	dobujack.catalog	54,800	3,620
13	welldone_catalogue	14,100	447
14	gozealcatalog	109,000	207
15	microzide.catalog	11,000	207
16	insurgentstore	213,000	200
17	hammerstout.catalog	40,000	133
18	unkl347hq	180,000	342
19	lawless.store	136,000	1,146
20	popculinewords	65,400	309
21	shroudcatchlog	390	51

batch size 64. The learning rates for the pre-trained DNN layers and the last layer are set to 10^{-5} and 10^{-4}, respectively.

4.3 Training

We use one GTX1080 GPU to train the image translation module for 200 epochs, with learning rate of 0.0001, and Adam optimization with $(\beta_1, \beta_2) = (0.5, 0.999)$. Two modules are trained separately. We also fine-tune Intrinsic and DMIT based modules on the Instagram dataset to improve the popularity performance. The code and dataset are available on https://github.com/baharehmn/thesis.

4.4 Evaluation metrics

Evaluation of images quality is a complicated task, especially in practical aspects [5]. Authors in [5] propose a method that obtains the saliency map and its complementary region in the original image. They consider two terms, namely penalty, and compensation, to make the assessment more realistic. However, to perform a fair comparison with the base methods, MUNIT [22], DRIT [30], and SignleGAN [49], we use FID and LPIPS as our evaluation metrics [20, 50]. The lower the FID, the quality and diversity of the generated images are better. Also, the higher the LPIPS score is desired and indicates that the generated images are diverse.

The inception distance Fréchet (FID) FID is introduced as a metric for evaluating the quality of images generated by GANs in 2017 [20]. An inception network [44] is first used to extract the middle layer features to calculate the FID. According to the obtained features, we model the data distribution using the multi-variate Gaussian distance distribution with mean μ and covariance Σ. The FID between the actual X image and the generated G image is calculated as follows:

$$FID(x, g) = \|\mu_x - \mu_g\|_2^2 + Tr\left(\sum_x + \sum_g - 2\left(\sum_x \sum_g\right)^{\frac{1}{2}}\right) \tag{13}$$

Where Tr summarizes diagonal elements of the matrix. FID is sensitive to mode collapse, and therefore, if the model generates only one image for each class, the FID score increases. The lower the FID, the higher is the quality and diversity of the generated images. The similarity between the generated and real samples is obtained by measuring the distance between their distributions.

Learned perceptual image patch similarity (LPIPS) LPIPS was introduced in 2018 to examine the diversity of generated images [50]. LPIPS is calculated from the same inputs using the weight distance between the random translation results in-depth features. The higher the LPIPS score, the better, indicating that the generated images are diverse. The lower the LIPIPS score for the model, the generated images are similar and have less diversity.

4.5 Results

In this section, we evaluate the quality of images generated by the proposed model with FID [20] and LPIPS [50] evaluation metrics and compare the results with the output of images generated by the MUNIT [22], DRIT [30] and SignleGAN [49]. We also discuss the popularity of the images generated by the DMIT based I2I Translation Module using the intrinsic evaluation metric. We train the I2I Translation Module with two modes using two types of data (color images and edge images). Since we use MUNIT as the base model [22], we consider the same two training modes (using two data) from the MUNIT model. The compared methods are also trained according to these two modes. The evaluation starts right after the training of models.

To evaluate the base models [22], we consider three scenarios as follows:

Scenario I; We use the input content image, the edge image A, and also the input style and the color image B. Now the model converts the C image based on the style and content of the input. In scenario I, the C image is the result of translation $A \rightarrow B$ from the content Image A to the style and the color from image B. The sampling procedure is shown in Fig. 5.

Scenario II; For translation $B \rightarrow A$, we set the input of the MUNIT as a color image B, the content image, and the style of the MUNIT network as an edge image A. Now, the MUNIT model generates the C image based on these two input images. The model outputs (two samples) are shownin Figs. 6 and 7.

Scenario III; We give both style and content inputs of the MUNIT network from the color images, and then translate $A \rightarrow B$ once and translate $B \rightarrow A$ again; the results are shown in Fig. 8:

Fig. 5 MUNIT [22] Image Translation procedure - Scenario I

Fig. 6 MUNIT [22] Image Translation procedure - Scenario II - sample I

Fig. 7 MUNIT [22] Image Translation procedure - Scenario II - Sample II

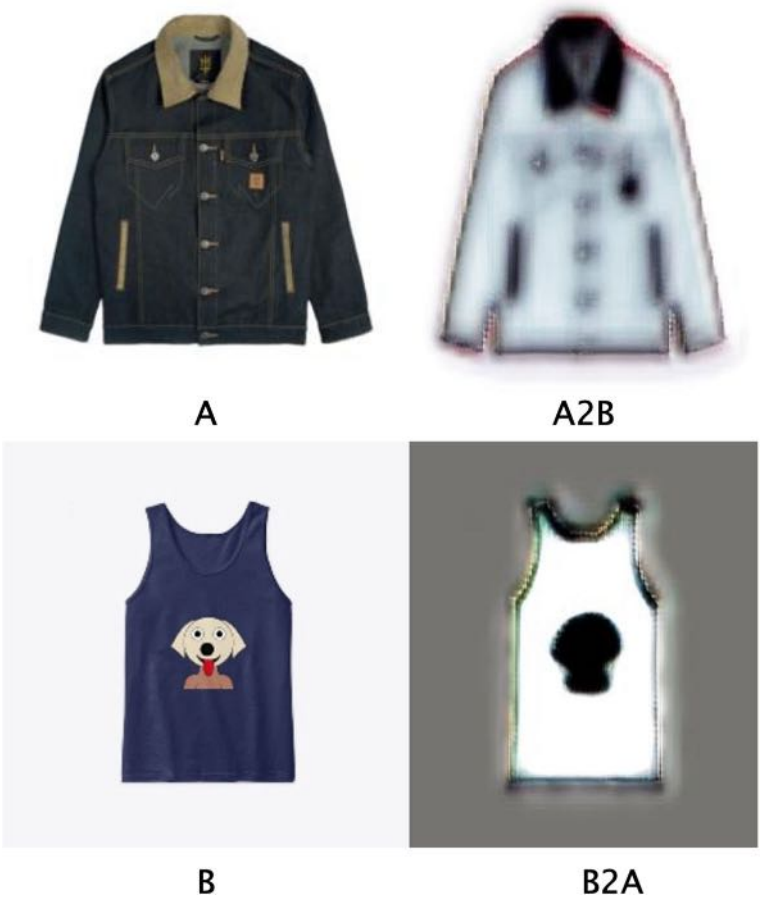


Fig. 8 MUNIT Image Translation procedure - Scenario III

Now, we evaluate the proposed method. We consider two following cases;

First Case: we set the edge image A and the color image B as inputs of the network. The model's output is ten images, and the network itself performs four translations: $A \rightarrow B$, $B \rightarrow A$, $B \rightarrow B$, $A \rightarrow A$. The results (two samples) are shown in Figs. 9 and 10,

The translation $A \rightarrow B$, the output image $encA \rightarrow B$ of that includes the shape of image A and the color of the image B.

The translation $B \rightarrow A$, the output image $encB \rightarrow A$ that includes the shape or the content from image B and the color or the style from image A, which is black and white.

Translation $A \rightarrow A$ consists of 4 images in which image A is generated in 4 different random colors.

The translation $B \rightarrow B$ output consists of 4 images in which image B is generated in 4 different random colors.

The outputs of the translations $A \rightarrow A$ and $B \rightarrow B$ confirm the diversity of the proposed model's output.

Fig. 9 Proposed image-to-image Translation procedure- First Case- sample I

Second Case; We use two-color images A and B as the model's input. The model's output contains ten images that are the result of $A \rightarrow B$, $B \rightarrow A$, $B \rightarrow B$, $A \rightarrow A$ translations. The results (four samples) are shown in Figs. 11, 12, 13 and 14

By reviewing the generated outputs in both cases, it is shown that the quality of the images generated in the second case is better compared with the quality of the images generated in the first case.

According to the results obtained by our proposed model, MUNIT [22], DRIT [30], and SignleGAN [49], we evaluate the quality and diversity of images generated by both models in terms of FID [20] and LPIPS [50] metrics. The results obtained by applying FID and LPIPS evaluation metrics on the the proposed method, MUNIT [22], DRIT [30], and SignleGAN [49], are summarized in Table 2.

For the comparison in Table 2, we consider both training modes in the first and second cases. In the first case, we compare the images generated based on the translation of the edge

Fig. 10 Proposed image-to-image Translation procedure- First Case- sample II

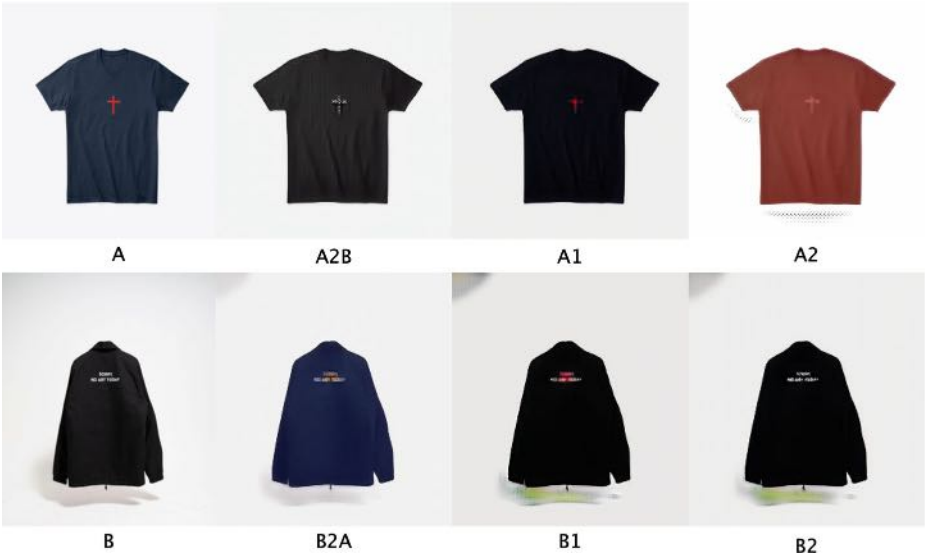


Fig. 11 Proposed image-to-image Translation procedure- Second Case- Sample I

image to MUNIT [22], DRIT [30], SingleGAN [49], and the proposed method in terms of quality and diversity. In the second case, we compare the quality and diversity of generated images based on the translation of color images to color images by the MUNIT [22], DRIT [30], SingleGAN [49], and proposed method. In other words, in the first case, we use input A as an edge photo for both models and input B as a color image, and translation is performed

Fig. 12 Proposed image-to-image Translation procedure- Second Case- Sample II

Fig. 13 Proposed image-to-image Translation procedure- Second Case- Sample III

for $A2B$ and $B2A$. In the second case, we consider A and B as color image inputs for both models, and we perform the translation for $A2B$ and $B2A$.

Results show that the images generated by the proposed method in the second case have a higher quality when both the input images are color, and the content and style are selected from color images instead of edge images.

We use the FID score to evaluate the quality of the generated images. As shown in Fig. 15, the FID score for the MUNIT model [22], DRIT [30], and SingleGAN [49] is higher than the proposed method, indicating that the quality of images generated by our proposed model is much better compared with the MUNIT [22], DRIT [30] and SingleGAN [49]. In color-to-color translation (second case), the proposed model shows around 80% improvement compared with the MUNIT [22] and around 60% improvement compared with DRIT [30], and SingleGAN [49] in terms of FID score. In Edge-to-color translation (first case), the improvement is around 50% compared with the MUNIT [22], 26% compared with the DRIT [30] and around 30% compared with the SingleGAN [49] in terms of FID score. One of the reasons is that, unlike the MUNIT [22], DRIT [30], and SingleGAN [49], the proposed DMIT based model performs multi-dimension and multi-domain translation while the compared methods performs the one-to-many mapping between the domains of two images. Unlike the proposed method, the MUNIT model [22], DRIT [30], and SingleGAN [49] do not use an integrated and unified generator to generate images.

The LPIPS score shows the diversity of the generated images. Higher values of the LPIPS score shows more diversity on the output images. As can be seen in Fig. 16, the MUNIT [22], DRIT [30], and SingleGAN [49] have lower LPIPS score compared to the proposed DMIT based method. It means that the images generated by the proposed method

Fig. 14 Proposed image-to-image Translation procedure- Second Case- Sample IV

have more diversity compared with the MUNIT [22], DRIT [30], and SingleGAN [49]. The LPIPS shows around 50% improvements on the color-to-color translation compared with the MUNIT [22] and around 50% improvement on edge-to-color image translation compared with MUNIT [22], DRIT [30], and SingleGAN [49]. The performance of the proposed DMIT based method is almost similar in color-to-color translation compared with the DRIT

Table 2 Comparison of proposed method with MUNIT [22] and DRIT [30] and SingleGAN [49] in terms of FID and LPIPS

	Edge to Color		Color to Color	
	LPIPS	FID	LPIPS	FID
MUNIT [22]	0.33	109	0.29	261
DRIT [30]	0.23	69	0.45	128
SingleGAN [49]	0.28	72	0.57	132
Proposed Method (DMIT based)	0.62	51.04	0.58	52.05

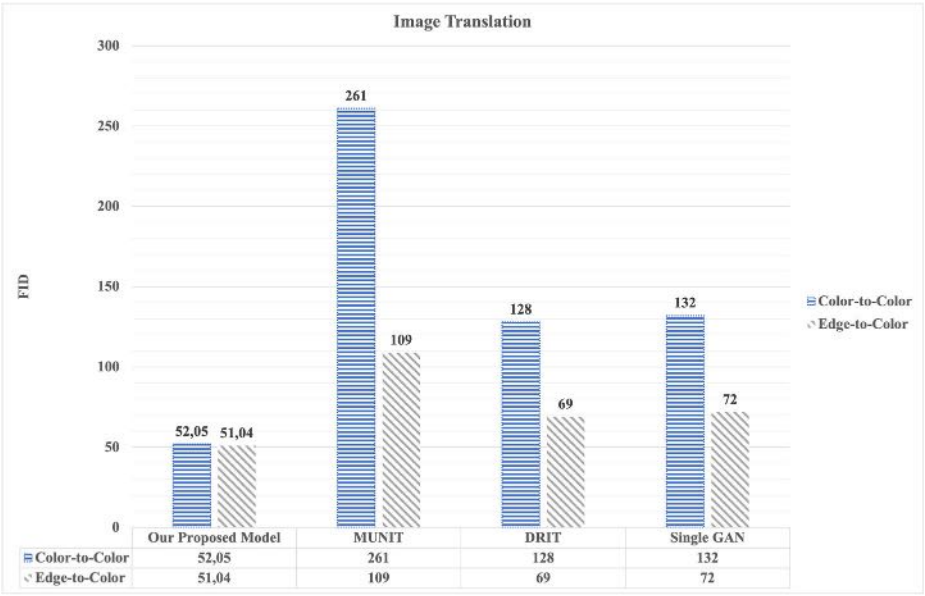


Fig. 15 Comparison of the quality of the generated images by proposed method with the MUNIT [22], DRIT [30] and SingleGAN [49] in terms of FID score

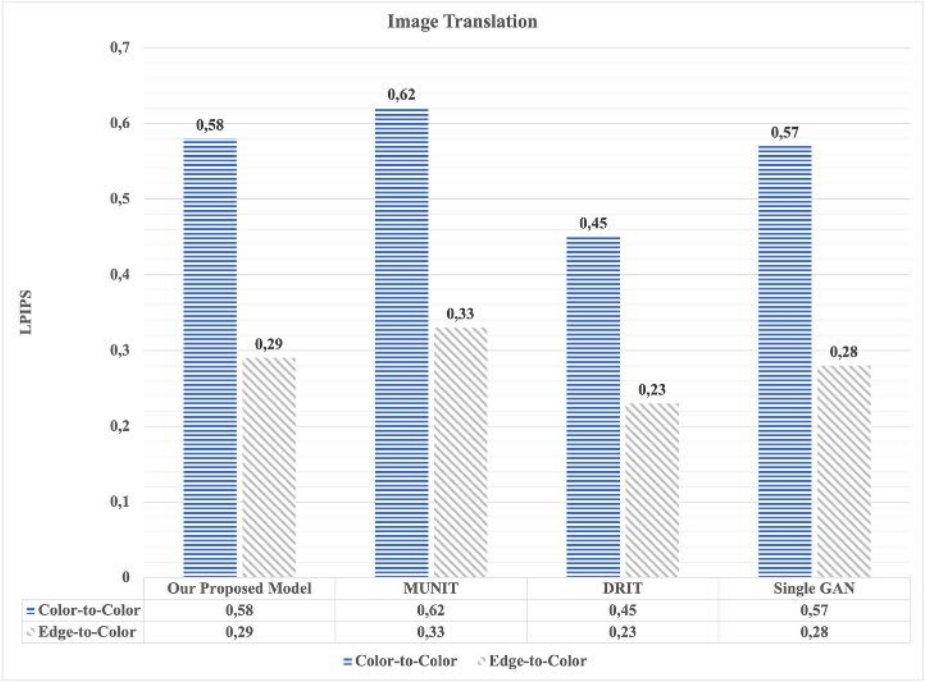


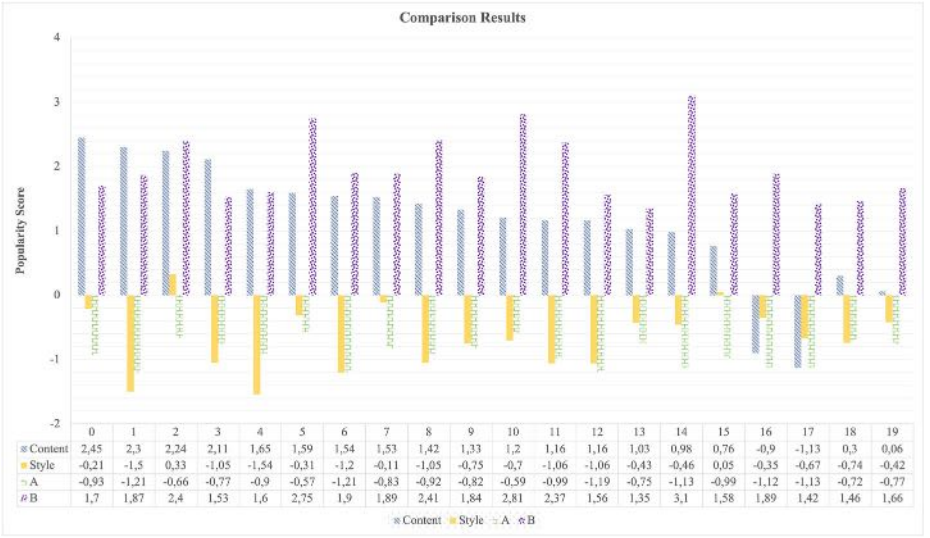
Fig. 16 Comparison of the quality of the generated images by proposed method with the MUNIT [22], DRIT [30] and SingleGAN [49] in terms of LPIPS

[30], and SingleGAN [49]. This is because, unlike the MUNIT [22], DRIT [30], and SingleGAN [49], the proposed DMIT based method performs multi-domain translation as well as using a unified generator.

Next, we evaluate the popularity of images generated by the proposed method. To perform this evaluation, we use the intrinsic evaluation metric [12]. The intrinsic evaluation metric is usually used to evaluate the intrinsic popularity of real images; however, we apply this metric to synthetic images generated by the proposed DMIT based method. We show that by applying the intrinsic network to our dataset, we could predict the popularity of the generated images and the popularity of the real images.

To perform the popularity evaluation, we use the popularity module to classify our collected dataset into popular and non-popular categories. These two categories are the inputs for the image-to-image translation module. We train the image generation module (second module) to learn the content and the style of popular and non-popular images and define different loss functions for content and style. The generated images are given to the popularity module for evaluating the popularity of the generated images by using the intrinsic evaluation metric. The results show that after training our model, the generated images with the same content are popular as the input images, and their style is similar to the popular input images. In other words, we can feed less popular or unpopular images as input to our model and then translate the style or the content of these images by using the style or the content of the popular images and finally generate images with the popular style or the popular content.

The score obtained through the intrinsic metric is proportional to the number of likes received by images and can distinguish between popular and non-popular images. As shown in Fig. 17, 40 test color images have been selected in categories A and B, and the intrinsic



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Content	2,45	2,3	2,24	2,11	1,65	1,59	1,54	1,53	1,42	1,33	1,2	1,16	1,16	1,03	0,98	0,76	-0,9	-1,13	0,3	0,06
Style	-0,21	-1,5	0,33	-1,05	-1,54	-0,31	-1,2	-0,11	-1,05	-0,75	-0,7	-1,06	-1,06	-0,43	-0,46	0,05	-0,35	-0,67	-0,74	-0,42
A	-0,93	-1,21	-0,66	-0,77	-0,9	-0,57	-1,21	-0,83	-0,92	-0,82	-0,59	-0,99	-1,19	-0,75	-1,13	-0,99	-1,12	-1,13	-0,72	-0,77
B	1,7	1,87	2,4	1,53	1,6	2,75	1,9	1,89	2,41	1,84	2,81	2,37	1,56	1,35	3,1	1,58	1,89	1,42	1,46	1,66

Fig. 17 The Popularity Score of the selected images by intrinsic evaluation metric [12] ranges from -2 to 3: the more the score, the more popular the image. The images in category A are unpopular, and images in category B are popular. The image-to-image translation is performed on content or style (based on image B), and the post-translation popularity of image A is presented in content and style rows

popularity score is used to calculate their popularity. The image popularity score is in the range of -2 to 3. In Fig. 17, **content** rows show the popularity score after translating unpopular image A by the content of the popular image B. The **style** rows also show the popularity score after translating the unpopular image A based on the style of the popular image, B.

We feed these selected images presented in Fig. 17 to the proposed method by using $A \rightarrow B$ and $B \rightarrow A$ translation. As can be seen, the unpopular image A has a higher popularity score after translation by $B \rightarrow A$ and resembles the shape (content) of the popular image B. Also, when unpopular image A is translated to the color version by $A \rightarrow B$, it receives the color of the popular image B, and its popularity score increases.

Comparing the popularity score of images generated from content translation versus the popularity score of images generated from style translation shows that translating an unpopular image into a popular image shape (content) has a more significant effect on increasing the image's popularity score. The generated images by this method will be more prevalent in the real world from the users' point of view. Figures 18, 19 and 20 show the popularity score for some samples of generated images by our model. The results depicted in Figs. 18, 19, and 20 show that the images generated by our model are more popular (have a better popularity score) compared with the original image due to changes we perform on the style or the content of images.

By reviewing the images' popularity score, we found out that the images generated using a content translation have a higher popularity score compared with the images generated by the style translation. Evaluation results show that 70% of images generated with content translator have a score greater than 1. This indicates that the role of content in generating popular images is much more significant compared with the style-based popular image generation.

Fig. 18 Popularity Score of generated images by Proposed Model-sample I

Fig. 19 Popularity Score of generated images by Proposed Model-sample II

Fig. 20 Popularity Score of generated images by Proposed Model-sample III

5 Conclusions

In this paper, we studied image-to-image translation for industrial use. The existing research works do not use popularity scores in image generation. Therefore, we proposed a new method to generate synthetic images in the fashion industry (focusing on clothes) that generate images according to users' opinions. In our research, we used a combination of generative adversarial networks (GAN), deep learning, and users' opinions. We used the ResNet50 neural network as the default deep neural network by modifying its architecture to evaluate the collected dataset. In the ResNet50 model, the last layer is replaced with a fully connected layer. The input images' size is set to 256x256, and we used the Adam function for optimization with a batch size of 64. The multi-dimensional translation and multi-domain translation on our proposed model consists of two modules, one is a popularity module that calculates image popularity by data classification, and the other one is the translation module (image-to-image generation) that converts unpopular images into popular ones. Our proposed model decreases the design time and cost of while achieving creativity and diversity. To calculate the images popularity, we first generated an image database that includes data popularity-discriminable image pairs (PDIPs) and applied the deep neural network computational model to the dataset. Unlike previous methods that use different discriminators for different domains, this DMIT based module uses a unified conditional discriminator for different domains. Our dataset consists of images along with the number of likes, download time, post address, user ID, content type, image post time, caption including the hashtag, and the number of comments for each photo. We collected around 130,000 images from Instagram, including the needed metadata. In the experiments, we evaluated the quality and diversity of the generated images in different scenarios and performed a comparative analysis. In color-to-color and edge-to-color translation, the proposed model showed at-least 60% and 25% improvement in terms of FID score, respectively. In terms of the LPIPS score, the proposed method showed around 50% improvement and showed that the images generated by the proposed method have more diversity. The evaluation results showed that it is possible to turn the unpopular input image into popular images by translating the content and style. Measuring the popularity score of images generated from content translation versus the popularity score of images generated from style translation showed that translating an unpopular image into a popular image shape (content) had significant effect on increasing the image's popularity score. There is still much to do on improving the accuracy and the performance of the proposed method and its evaluation on different categories that remain as the future works of this research.

Declarations

The author declare that there are no conflicts of interest regarding the publication of this paper.

References

1. Achanta SDM, Karthikeyan T, Vinoth Kanna R (2021) Wearable sensor based acoustic gait analysis using phase transition-based optimization algorithm on iot. Int J Speech Technol, pp 1–11
2. Achanta SDM, Karthikeyan T, Vinothkanna R (2019) A novel hidden markov model-based adaptive dynamic time warping (hmdtw) gait analysis for identifying physically challenged persons. Soft Comput 23(18):8359–8366

3. Achanta SDM, Karthikeyan T et al (2019) A wireless iot system towards gait detection technique using fsr sensor and wearable iot devices. Int J Intell Unmanned Syst
4. Alec R, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434
5. Amirkhani D, Bastanfard A (2021) An objective method to evaluate exemplar-based inpainted images quality using jaccard index. Multimed Tools Appl 80(17):26199–26212
6. Antreas A, Storkey A, Edwards H (2017) Data augmentation generative adversarial networks. arXiv:1711.04340
7. Bai J, Chen R, Liu M (2020) Feature-attention module for context-aware image-to-image translation. Vis Comput 36(10):2145–2159
8. Chai C, Liao J, Zou N, Sun L (2018) A one-to-many conditional generative adversarial network framework for multiple image-to-image translations. Multimed Tools Appl 77(17):22339–22366
9. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016) Infogan: interpretable representation learning by information maximizing generative adversarial nets. arXiv:1606.03657
10. Cheng G, Sun X, Li K, Guo L, Han J (2021) Perturbation-seeking generative adversarial networks: a defense framework for remote sensing image scene classification. IEEE Trans Geosci Remote Sensing
11. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Stargan JC (2018) Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
12. Ding K, Ma K, Wang S (2019) Intrinsic image popularity assessment. In: Proceedings of the 27th ACM international conference on multimedia, pp 1979–1987
13. 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
14. 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
15. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial networks. arXiv:1406.2661
16. Gothwal R, Gupta S, Gupta D, Dahiya AK (2014) Color image segmentation algorithm based on rgb channels. In: Proceedings of 3rd international conference on reliability, infocom technologies and optimization, pp 1–5
17. Hajarian M, Bastanfard A, Mohammadzadeh J, Khalilian M (2017) Introducing fuzzy like in social networks and its effects on advertising profits and human behavior. Comput Hum Behav 77:282–293
18. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
19. Hessel J, Lee L, Mimno D (2017) Cats and captions vs. creators and the clock: comparing multimodal content to context in predicting relative popularity. In: Proceedings of the 26th international conference on world wide web, pp 927–936
20. Heusel M, Ramsauer H, Unterthiner T (2017) Bernhard nessler, and sepp hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. arXiv:1706.08500
21. Hsu C-C, Hwang H-T, Wu Y-C, Tsao Y, Wang H-M (2017) Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. arXiv:1704.00849
22. Huang X, Liu M-Y, Belongie S, Kautz J (2018) Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV), pp 172–189
23. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
24. Jun-Yan Zhu, Zhang R, Pathak D, Trevor D, Alexei AE, Wang O, Shechtman E (2017) Toward multimodal image-to-image translation. arXiv:1711.11586
25. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196
26. Khosla A, Sarma AD, Hamid R (2014) What makes an image popular? In: Proceedings of the 23rd international conference on World wide web, pp 867–876
27. Kingma DP, Adam JB (2014) A method for stochastic optimization. arXiv:1412.6980
28. Kingma DP, Welling M (2014) Stochastic gradient vb and the variational auto-encoder. In: Second international conference on learning representations, ICLR, vol 19
29. Kupyn O, Budzan V, Mykhailych M, Mishkin D, Deblurgan JM (2018) Blind motion deblurring using conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8183–8192

30. Lee H-Y, Tseng H-Y, Huang J-B, Singh M, Yang M-H (2018) Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV), pp 35–51
31. Lin K, Li D, He X, Zhang Z, Sun M-T (2017) Adversarial ranking for language generation. arXiv:1705.11001
32. Liu M-Y, Breuel T, Jan Kautz (2017) Unsupervised image-to-image translation networks. arXiv:1703.00848
33. Liu M-Y, Breuel T, Kautz J (2017) Unsupervised image-to-image translation networks. arXiv:1703.00848
34. Liu Z, Gao F, Wang Y (2019) A generative adversarial network for ai-aided chair design. In: IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 486–490
35. Liu M-Y, Huang X, Yu J, Wang T-C, Mallya A (2020) Generative adversarial networks for image and video synthesis: algorithms and applications. arXiv:2008.02793
36. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv:1411.1784
37. Murthy ASD, Karthikeyan T, Vinoth Kanna R (2021) Gait-based person fall prediction using deep learning approach. Soft Comput, pp 1–9
38. Na L, Zheng Z, Zhang S, Zhibin Y, Zheng H, Zheng B (2018) The synthesis of unpaired underwater images using a multistyle generative adversarial network. IEEE Access 6:54241–54257
39. Qian X, Xi C, Cheng G, Yao X, Jiang L (2021) Two-stream encoder gan with progressive training for co-saliency detection. IEEE Signal Process Lett 28:180–184
40. Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and variational inference in deep latent gaussian models. In: International conference on machine learning. Citeseer, vol 2, p 2
41. Richardson E, Alaluf Y, Or P, Nitzan Y, Azar Y, Shapiro S, Cohen-Or D (2021) Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2287–2296
42. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. arXiv:1606.03498
43. Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. Adv Neural Inform Process Syst 28:3483–3491
44. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2818–2826
45. Tian Y, Peng X, Zhao L, Zhang S, Metaxas DN (2018) Cr-gan: learning complete representations for multi-view generation. arXiv:1806.11191
46. Wang C, Chang Xu, Wang C, Tao D (2018) Perceptual adversarial networks for image-to-image transformation. IEEE Trans Image Process 27(8):4066–4079
47. Wang W, Zhou W, Bao J, Chen D, Li H (2021) Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. arXiv:2108.04547
48. Xiaoming Y, Chen Y, Li T, Liu S, Li G (2019) Multi-mapping image-to-image translation via learning disentanglement. arXiv:1909.07877
49. Yu X, Cai X, Ying Z, Li T, Li G (2018) Singlegan: image-to-image translation by a single-generator network using multiple generative adversarial learning. In: Asian conference on computer vision. Springer, pp 341–356
50. Zhang R, Isola P, Efros AA, 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
51. Zhao Y, Zheng Z, Wang C, Zhaorui G, Min F, Zhibin Y, Zheng H, Wang N, Zheng B (2020) Fine-grained facial image-to-image translation with an attention based pipeline generative adversarial framework. Multimed Tools Appl, pp 1–20
52. Zhu J-Y, Zhang R, Pathak D, Darrell T, Efros AA (2017) Oliver wang, and eli shechtman. Toward multimodal image-to-image translation. arXiv:1711.11586