**StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation**

Yunjey Choi1,2 Minje Choi1,2 Munyoung Kim2,3 Jung-Woo Ha2 Sunghun Kim2,4 Jaegul Choo1,2 1 Korea University 2 Clova AI Research, NAVER Corp. 3 The College of New Jersey 4 Hong Kong University of Science & Technology

**Abstract**

Recent studies have shown remarkable success in imageto-image translation for two domains. However, existing approaches have limited scalability and robustness in handling more than two domains, since different models should be built independently for every pair of image domains. To address this limitation, we propose StarGAN, a novel and scalable approach that can perform image-to-image translations for multiple domains using only a single model. Such a unified model architecture of StarGAN allows simultaneous training of multiple datasets with different domains within a single network. This leads to StarGAN’s superior quality of translated images compared to existing models as well as the novel capability of flexibly translating an input image to any desired target domain. We empirically demonstrate the effectiveness of our approach on a facial attribute transfer and a facial expression synthesis tasks.

최근 연구는 두 영역에 대한 이미지 대 이미지 번역에서 놀라운 성공을 보여주었습니다. 그러나 기존 접근 방식은 모든 이미지 도메인 쌍에 대해 서로 다른 모델을 독립적으로 구축해야 하므로 2개 이상의 도메인을 처리하는 데 확장성과 견고성이 제한적입니다. 이러한 한계를 해결하기 위해 단일 모델을 사용하여 여러 도메인에 대한 이미지 대 이미지 번역을 수행할 수 있는 새롭고 확장 가능한 접근 방식인 StarGAN을 제안합니다. 이러한 StarGAN의 통합 모델 아키텍처를 통해 단일 네트워크 내에서 서로 다른 도메인을 가진 여러 데이터 세트를 동시에 훈련할 수 있습니다. 이는 StarGAN의 기존 모델에 비해 우수한 번역 이미지 품질과 입력 이미지를 원하는 대상 도메인으로 유연하게 번역하는 새로운 기능으로 이어집니다. 우리는 얼굴 속성 전달 및 표정 합성 작업에 대한 접근 방식의 효과를 경험적으로 보여줍니다.

**1. Introduction**

텍스트, 가장, 그룹, 사람이(가) 표시된 사진

자동 생성된 설명

Figure 1. Multi-domain image-to-image translation results on the CelebA dataset via transferring knowledge learned from the RaFD dataset. The first and sixth columns show input images while the remaining columns are images generated by StarGAN. Note that the images are generated by a single generator network, and facial expression labels such as angry, happy, and fearful are from RaFD, not CelebA.

그림 1. RaFD 데이터 세트에서 학습한 지식 전달을 통한 CelebA 데이터 세트의 다중 도메인 이미지-이미지 변환 결과. 첫 번째와 여섯 번째 열은 입력 이미지이고 나머지 열은 StarGAN에서 생성된 이미지입니다. 이미지는 단일 생성기 네트워크에서 생성되며 분노, 행복, 두려움과 같은 표정 레이블은 CelebA가 아닌 RaFD에서 가져온 것입니다.

The task of image-to-image translation is to change a particular aspect of a given image to another, e.g., changing the facial expression of a person from smiling to frowning (see Fig. 1). This task has experienced significant improvements following the introduction of generative adversarial networks (GANs), with results ranging from changing hair color [9], reconstructing photos from edge maps [7], and changing the seasons of scenery images [33].이미지를 이미지로 변환하는 작업은 주어진 이미지의 특정 측면을 다른 것으로 변경하는 것입니다(예: 웃는 얼굴 표정에서 찡그린 얼굴 표정으로 변경)(그림 1 참조). 이 작업은 GAN(Generative Adversarial Networks)의 도입 후 머리 색깔 변경[9], 에지 맵에서 사진 재구성[7], 풍경 이미지의 계절 변경[33]에 이르는 결과로 상당한 개선을 경험했습니다.

Given training data from two different domains, these models learn to translate images from one domain to the other. We denote the terms attribute as a meaningful feature inherent in an image such as hair color, gender or age, and attribute value as a particular value of an attribute, e.g., black/blond/brown for hair color or male/female for gender. We further denote domain as a set of images sharing the same attribute value. For example, images of women can represent one domain while those of men represent another.

두 개의 서로 다른 도메인의 훈련 데이터가 주어지면 이 모델은 한 도메인에서 다른 도메인으로 이미지를 번역하는 방법을 배웁니다. 우리는 속성이라는 용어를 머리 색깔, 성별 또는 나이와 같은 이미지에 내재된 의미 있는 특징으로, 속성 값을 속성의 특정 값으로 표시합니다(예: 머리 색깔은 검정/금발/갈색, 성별은 남성/여성). 동일한 속성 값을 공유하는 이미지 집합으로 도메인을 추가로 나타냅니다. 예를 들어, 여성의 이미지는 한 영역을 나타내고 남성의 이미지는 다른 영역을 나타낼 수 있습니다.

Several image datasets come with a number of labeled attributes. For instance, the CelebA[19] dataset contains 40 labels related to facial attributes such as hair color, gender, and age, and the RaFD [13] dataset has 8 labels for facial expressions such as ‘happy’, ‘angry’ and ‘sad’. These settings enable us to perform more interesting tasks, namely multi-domain image-to-image translation, where we change images according to attributes from multiple domains. The first five columns in Fig. 1 show how a CelebA image can be translated according to any of the four domains, ‘blond hair’, ‘gender’, ‘aged’, and ‘pale skin’. We can further extend to training multiple domains from different datasets, such as jointly training CelebA and RaFD images to change a CelebA image’s facial expression using features learned by training on RaFD, as in the rightmost columns of Fig. 1.

여러 이미지 데이터 세트에는 여러 레이블이 지정된 속성이 있습니다. 예를 들어 CelebA[19] 데이터 세트에는 머리색, 성별, 나이와 같은 얼굴 속성과 관련된 40개의 레이블이 포함되어 있고 RaFD[13] 데이터 세트에는 'happy', 'angry' 및 '와 같은 얼굴 표정에 대한 8개의 레이블이 있습니다. 슬퍼'. 이러한 설정을 통해 여러 도메인의 속성에 따라 이미지를 변경하는 다중 도메인 이미지 대 이미지 번역과 같은 보다 흥미로운 작업을 수행할 수 있습니다. 그림 1의 처음 5개 열은 CelebA 이미지가 '금발 머리', '성별', '노화', '창백한 피부'의 네 가지 영역 중 하나에 따라 어떻게 번역될 수 있는지 보여줍니다. 그림 1의 맨 오른쪽 열에서와 같이 RaFD에서 학습한 기능을 사용하여 CelebA 이미지의 얼굴 표정을 변경하기 위해 CelebA 및 RaFD 이미지를 공동으로 훈련하는 것과 같이 서로 다른 데이터 세트에서 여러 도메인을 훈련하도록 확장할 수 있습니다.

However, existing models are both inefficient and ineffective in such multi-domain image translation tasks. Their inefficiency results from the fact that in order to learn all mappings among k domains, k(k−1) generators have to be trained. Fig. 2 (a) illustrates how twelve distinct generator networks have to be trained to translate images among four different domains. Meanwhile, they are ineffective that even though there exist global features that can be learned from images of all domains such as face shapes, each generator cannot fully utilize the entire training data and only can learn from two domains out of k. Failure to fully utilize training data is likely to limit the quality of generated images. Furthermore, they are incapable of jointly training domains from different datasets because each dataset is partially labeled, which we further discuss in Section 3.2.

그러나 기존 모델은 이러한 다중 도메인 이미지 번역 작업에서 비효율적이며 비효율적입니다. k 도메인 간의 모든 매핑을 학습하려면 k(k−1) 생성기를 훈련해야 하기 때문에 비효율적입니다. 그림 2(a)는 4개의 다른 도메인 사이에서 이미지를 번역하기 위해 12개의 별개의 생성기 네트워크가 어떻게 훈련되어야 하는지를 보여줍니다. 한편, 얼굴형과 같은 모든 영역의 이미지에서 학습할 수 있는 전역적 특징이 존재하더라도 각 Generator는 전체 학습 데이터를 충분히 활용하지 못하고 k 중 2개 영역에서만 학습할 수 있다는 비효율적입니다. 훈련 데이터를 충분히 활용하지 못하면 생성된 이미지의 품질이 제한될 수 있습니다. 또한 각 데이터 세트에 부분적으로 레이블이 지정되어 있기 때문에 서로 다른 데이터 세트의 도메인을 공동으로 훈련할 수 없습니다. 이에 대해서는 섹션 3.2에서 자세히 설명합니다.

As a solution to such problems we propose StarGAN, a novel and scalable approach capable of learning mappings among multiple domains. As demonstrated in Fig. 2 (b), our model takes in training data of multiple domains, and learns the mappings between all available domains using only a single generator. The idea is simple. Instead of learning a fixed translation (e.g., black-to-blond hair), our generator takes in as inputs both image and domain information, and learns to flexibly translate the image into the corresponding domain. We use a label (e.g., binary or one-hot vector) to represent domain information. During training, we randomly generate a target domain label and train the model to flexibly translate an input image into the target domain. By doing so, we can control the domain label and translate the image into any desired domain at testing phase.

이러한 문제에 대한 솔루션으로 여러 도메인 간의 매핑을 학습할 수 있는 새롭고 확장 가능한 접근 방식인 StarGAN을 제안합니다. 그림 2(b)에서 볼 수 있듯이 우리 모델은 여러 도메인의 학습 데이터를 가져와 단일 생성기를 사용하여 사용 가능한 모든 도메인 간의 매핑을 학습합니다. 아이디어는 간단합니다. 고정 번역(예: 흑발에서 금발)을 학습하는 대신 생성기는 이미지와 도메인 정보를 모두 입력으로 사용하고 이미지를 해당 도메인으로 유연하게 번역하는 방법을 학습합니다. 도메인 정보를 나타내기 위해 레이블(예: 이진 또는 원-핫 벡터)을 사용합니다. 훈련 중에 대상 도메인 레이블을 무작위로 생성하고 입력 이미지를 대상 도메인으로 유연하게 변환하도록 모델을 훈련합니다. 이렇게 하면 도메인 레이블을 제어하고 테스트 단계에서 이미지를 원하는 도메인으로 변환할 수 있습니다.

We also introduce a simple but effective approach that enables joint training between domains of different datasets by adding a mask vector to the domain label. Our proposed method ensures that the model can ignore unknown labels and focus on the label provided by a particular dataset. In this manner, our model can perform well on tasks such as synthesizing facial expressions of CelebA images using features learned from RaFD, as shown in the rightmost columns of Fig. 1. As far as our knowledge goes, our work is the first to successfully perform multi-domain image translation across different datasets.

또한 도메인 레이블에 마스크 벡터를 추가하여 서로 다른 데이터 세트의 도메인 간의 공동 훈련을 가능하게 하는 간단하지만 효과적인 접근 방식을 소개합니다. 우리가 제안한 방법은 모델이 알려지지 않은 레이블을 무시하고 특정 데이터 세트에서 제공하는 레이블에 집중할 수 있도록 합니다. 이러한 방식으로 우리 모델은 그림 1의 가장 오른쪽 열에 표시된 것처럼 RaFD에서 학습한 기능을 사용하여 CelebA 이미지의 얼굴 표정을 합성하는 것과 같은 작업을 잘 수행할 수 있습니다. 다양한 데이터세트에서 다중 도메인 이미지 번역을 수행합니다.

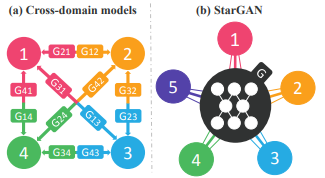


Figure 2. Comparison between cross-domain models and our proposed model, StarGAN. (a) To handle multiple domains, crossdomain models should be built for every pair of image domains. (b) StarGAN is capable of learning mappings among multiple domains using a single generator. The figure represents a star topology connecting multi-domains.

그림 2. 교차 도메인 모델과 제안 모델 StarGAN의 비교. (a) 여러 도메인을 처리하려면 모든 이미지 도메인 쌍에 대해 교차 도메인 모델을 구축해야 합니다. (b) StarGAN은 단일 생성기를 사용하여 여러 도메인 간의 매핑을 학습할 수 있습니다. 그림은 다중 도메인을 연결하는 스타 토폴로지를 나타냅니다.

Overall, our contributions are as follows:

전반적으로 우리의 기여는 다음과 같습니다:

• We propose StarGAN, a novel generative adversarial network that learns the mappings among multiple domains using only a single generator and a discriminator, training effectively from images of all domains.

• 단일 생성기와 판별기를 사용하여 여러 도메인 간의 매핑을 학습하고 모든 도메인의 이미지에서 효과적으로 훈련하는 새로운 생성적 적대 네트워크인 StarGAN을 제안합니다.

• We demonstrate how we can successfully learn multidomain image translation between multiple datasets by utilizing a mask vector method that enables StarGAN to control all available domain labels.

• StarGAN이 사용 가능한 모든 도메인 레이블을 제어할 수 있도록 하는 마스크 벡터 방법을 활용하여 여러 데이터 세트 간의 다중 도메인 이미지 변환을 성공적으로 학습하는 방법을 보여줍니다.

• We provide both qualitative and quantitative results on facial attribute transfer and facial expression synthesis tasks using StarGAN, showing its superiority over baseline models.

• StarGAN을 사용하여 얼굴 속성 전달 및 표정 합성 작업에 대한 정성적 및 정량적 결과를 모두 제공하여 기본 모델보다 우수함을 보여줍니다.

**2. Related Work**

Generative Adversarial Networks. Generative adversarial networks (GANs) [3] have shown remarkable results in various computer vision tasks such as image generation [6, 24, 32, 8], image translation [7, 9, 33], super-resolution imaging [14], and face image synthesis [10, 16, 26, 31]. A typical GAN model consists of two modules: a discriminator and a generator. The discriminator learns to distinguish between real and fake samples, while the generator learns to generate fake samples that are indistinguishable from real samples. Our approach also leverages the adversarial loss to make the generated images as realistic as possible.

Generative Adversarial Networks입니다. 생성적 적대 네트워크 (GANs)는 이미지 생성, 이미지 변환, 초해상도 이미징, 얼굴 이미지 합성과 같은 다양한 컴퓨터 비전 작업에서 놀라운 결과를 보였습니다. 일반적인 GAN 모델은 판별기와 발생기의 두 가지 모듈로 구성됩니다. 식별자는 실제 샘플과 가짜 샘플을 구별하는 방법을 배우고, 생성자는 실제 샘플과 구별할 수 없는 가짜 샘플을 생성하는 방법을 배웁니다. 또한 NAT의 접근 방식은 적대적 손실을 활용하여 생성된 이미지를 최대한 사실적으로 만듭니다.

Conditional GANs. GAN-based conditional image generation has also been actively studied. Prior studies have provided both the discriminator and generator with class information in order to generate samples conditioned on the class [20, 21, 22]. Other recent approaches focused on generating particular images highly relevant to a given text description [25, 30]. The idea of conditional image generation has also been successfully applied to domain transfer [9, 28], superresolution imaging[14], and photo editing [2, 27]. In this paper, we propose a scalable GAN framework that can flexibly steer the image translation to various target domains, by providing conditional domain information.

조건부 GAN. GAN 기반 조건부 이미지 생성도 활발히 연구되고 있다. 이전 연구에서는 클래스를 조건으로 하는 샘플을 생성하기 위해 판별자와 생성자 모두에 클래스 정보를 제공했습니다. 다른 최근 접근 방식은 주어진 텍스트 설명과 매우 관련성이 높은 특정 이미지를 생성하는 데 중점을 둡니다. 조건부 이미지 생성의 아이디어는 도메인 전송[9, 28], 초고해상도 이미징[14] 및 사진 편집[2, 27]에도 성공적으로 적용되었습니다. 본 논문에서는 조건부 도메인 정보를 제공하여 다양한 대상 도메인으로 이미지 번역을 유연하게 조정할 수 있는 확장 가능한 GAN 프레임워크를 제안합니다

텍스트, 스크린샷, 전자이(가) 표시된 사진

자동 생성된 설명

Figure 3. Overview of StarGAN, consisting of two modules, a discriminator D and a generator G. (a) D learns to distinguish between real and fake images and classify the real images to its corresponding domain. (b) G takes in as input both the image and target domain label and generates an fake image. The target domain label is spatially replicated and concatenated with the input image. (c) G tries to reconstruct the original image from the fake image given the original domain label. (d) G tries to generate images indistinguishable from real images and classifiable as target domain by D.그림 3. 판별기 D와 생성기 G의 두 가지 모듈로 구성된 StarGAN 개요 (a) D는 실제 이미지와 가짜 이미지를 구별하고 실제 이미지를 해당 도메인으로 분류하는 방법을 학습합니다. (b) G는 이미지와 대상 도메인 레이블을 모두 입력으로 받아 가짜 이미지를 생성합니다. 대상 도메인 레이블은 공간적으로 복제되고 입력 이미지와 연결됩니다. (c) G는 원본 도메인 레이블이 지정된 가짜 이미지에서 원본 이미지를 재구성하려고 시도합니다. (d) G는 실제 이미지와 구별할 수 없고 D에 의해 대상 도메인으로 분류될 수 있는 이미지를 생성하려고 합니다.

Image-to-Image Translation. Recent work have achieved impressive results in image-to-image translation [7, 9, 17, 33]. For instance, pix2pix [7] learns this task in a supervised manner using cGANs[20]. It combines an adversarial loss with a L1 loss, thus requires paired data samples. To alleviate the problem of obtaining data pairs, unpaired image-to-image translation frameworks [9, 17, 33] have been proposed. UNIT [17] combines variational autoencoders (VAEs) [12] with CoGAN [18], a GAN framework where two generators share weights to learn the joint distribution of images in cross domains. CycleGAN [33] and DiscoGAN [9] preserve key attributes between the input and the translated image by utilizing a cycle consistency loss. However, all these frameworks are only capable of learning the relations between two different domains at a time. Their approaches have limited scalability in handling multiple domains since different models should be trained for each pair of domains. Unlike the aforementioned approaches, our framework can learn the relations among multiple domains using only a single model.

이미지 대 이미지 번역. 최근 작업은 이미지 대 이미지 번역에서 인상적인 결과를 얻었습니다. 예를 들어 pix2pix[7]는 cGAN을 사용하여 감독 방식으로 이 작업을 학습합니다. 적대적 손실과 L1 손실을 결합하므로 쌍을 이루는 데이터 샘플이 필요합니다. 데이터 쌍을 얻는 문제를 완화하기 위해 짝을 이루지 않은 이미지-이미지 변환 프레임워크가 제안되었습니다. UNIT은 VAE(variational autoencoder)[12]와 CoGAN을 결합합니다. CoGAN은 두 개의 생성기가 가중치를 공유하여 교차 도메인에서 이미지의 공동 분포를 학습하는 GAN 프레임워크입니다. CycleGAN과 DiscoGAN은 주기 일관성 손실을 활용하여 입력과 번역된 이미지 사이의 주요 속성을 보존합니다. 그러나 이러한 모든 프레임워크는 한 번에 두 개의 서로 다른 도메인 간의 관계만 학습할 수 있습니다. 그들의 접근 방식은 각 도메인 쌍에 대해 서로 다른 모델을 교육해야 하기 때문에 여러 도메인을 처리하는 데 확장성이 제한적입니다. 앞서 언급한 접근 방식과 달리 우리의 프레임워크는 단일 모델을 사용하여 여러 도메인 간의 관계를 학습할 수 있습니다.

**3. Star Generative Adversarial Networks**

We first describe our proposed StarGAN, a framework to address multi-domain image-to-image translation within a single dataset. Then, we discuss how StarGAN incorporates multiple datasets containing different label sets to flexibly perform image translations using any of these labels.

먼저 단일 데이터 세트 내에서 다중 도메인 이미지 대 이미지 변환을 처리하기 위한 프레임워크인 제안된 StarGAN에 대해 설명합니다. 그런 다음 StarGAN이 다른 레이블 세트를 포함하는 여러 데이터 세트를 통합하여 이러한 레이블을 사용하여 이미지 번역을 유연하게 수행하는 방법을 논의합니다.

**3.1. Multi-Domain Image-to-Image Translation**

Our goal is to train a single generator G that learns mappings among multiple domains. To achieve this, we train G to translate an input image x into an output image y conditioned on the target domain label c, G(x, c) → y. We randomly generate the target domain label c so that G learns to flexibly translate the input image. We also introduce an auxiliary classifier [22] that allows a single discriminator to control multiple domains. That is, our discriminator produces probability distributions over both sources and domain labels, D : x → { }. Fig. 3 illustrates the training process of our proposed approach.

우리의 목표는 여러 도메인 간의 매핑을 학습하는 단일 제너레이터 G를 교육하는 것입니다. 이를 위해 우리는 G가 입력 이미지 x를 목표 도메인 레이블 c, G(x, c) → y에 따라 조건화된 출력 이미지 y로 변환하도록 훈련합니다. 우리는 G가 입력 이미지를 유연하게 변환하는 방법을 배우도록 목표 도메인 레이블 c를 무작위로 생성합니다. 또한 단일 식별자가 여러 도메인을 제어할 수 있도록 하는 보조 분류기 [22]를 도입합니다. 즉, 우리의 식별자는 소스 및 도메인 레이블 D : x → { }에 걸쳐 확률 분포를 생성합니다. 그림 3은 제안된 접근법의 교육 과정을 보여줍니다.

Adversarial Loss. To make the generated images indistinguishable from real images, we adopt an adversarial loss where G generates an image G(x, c) conditioned on boththe input image x and the target domain label c, while D tries to distinguish between real and fake images. In this paper, we refer to the term as a probability distribution over sources given by D. The generator G tries to minimize this objective, while the discriminator D tries to maximize it.적대적 손실. 생성된 이미지를 실제 이미지와 구별할 수 없도록 하기 위해 G는 입력 이미지 x와 대상 도메인 레이블 c 모두에 따라 이미지 G(x, c)를 생성하고 D는 실제 이미지와 가짜 이미지를 구별하려고 시도하는 적대적 손실을 채택합니다. 이 논문에서 우리는 라는 용어를 D에 의해 주어진 소스에 대한 확률 분포로 참조합니다. 생성자 G는 이 목적을 최소화하려고 시도하는 반면 판별자 D는 최대화를 시도합니다.



Domain Classification Loss. For a given input image x and a target domain label c, our goal is to translate x into an output image y, which is properly classified to the target domain c. To achieve this condition, we add an auxiliary classifier on top of D and impose the domain classification loss when optimizing both D and G. That is, we decompose the objective into two terms: a domain classification loss of real images used to optimize D, and a domain classification loss of fake images used to optimize G. In detail, the former is defined as

도메인 분류 손실. 주어진 입력 이미지 x와 대상 도메인 레이블 c에 대해 우리의 목표는 x를 대상 도메인 c로 적절하게 분류되는 출력 이미지 y로 변환하는 것입니다. 이 조건을 달성하기 위해 D 위에 보조 분류기를 추가하고 D와 G를 모두 최적화할 때 도메인 분류 손실을 부과합니다. 즉, 목적을 두 가지 용어로 분해합니다. D 최적화에 사용되는 실제 이미지의 도메인 분류 손실, 및 G 최적화에 사용된 가짜 이미지의 도메인 분류 손실. 구체적으로 전자는 다음과 같이 정의됩니다.



where the term represents a probability distribution over domain labels computed by D. By minimizing this objective, D learns to classify a real image x to its corresponding original domain c’. We assume that the input image and domain label pair (x, c’) is given by the training data. On the other hand, the loss function for the domain classification of fake images is defined as여기서 라는 용어는 D에 의해 계산된 도메인 레이블에 대한 확률 분포를 나타냅니다. 이 목적을 최소화함으로써 D는 실제 이미지 x를 해당하는 원래 도메인 c'로 분류하는 방법을 배웁니다. 입력 이미지와 도메인 레이블 쌍(x, c')이 학습 데이터에 의해 제공된다고 가정합니다. 한편, 가짜 이미지의 도메인 분류를 위한 손실 함수는 다음과 같이 정의됩니다.



In other words, G tries to minimize this objective to generate images that can be classified as the target domain c.

즉, G는 타겟 도메인 c로 분류할 수 있는 이미지를 생성하기 위해 이 목적을 최소화하려고 합니다.

Reconstruction Loss. By minimizing the adversarial and classification losses, G is trained to generate images that are realistic and classified to its correct target domain. However, minimizing the losses (Eqs. (1) and (3)) does not guarantee that translated images preserve the content of its input images while changing only the domain-related part of the inputs. To alleviate this problem, we apply a cycle consistency loss [9, 33] to the generator, defined as

재건 손실. 적대적 및 분류 손실을 최소화함으로써 G는 현실적이며 올바른 대상 도메인으로 분류되는 이미지를 생성하도록 훈련됩니다. 그러나 손실을 최소화(식 (1) 및 (3))한다고 해서 번역된 이미지가 입력의 도메인 관련 부분만 변경하면서 입력 이미지의 내용을 보존한다는 보장은 없습니다. 이 문제를 완화하기 위해 다음과 같이 정의된 발전기에 주기 일관성 손실을 적용합니다.



where G takes in the translated image G(x, c) and the original domain label c’ as input and tries to reconstruct the original image x. We adopt the L1 norm as our reconstruction loss. Note that we use a single generator twice, first to translate an original image into an image in the target domain and then to reconstruct the original image from the translated image.

여기서 G는 번역된 이미지 G(x, c)와 원본 도메인 레이블 c'를 입력으로 받아 원본 이미지 x를 재구성하려고 시도합니다. 우리는 재건 손실로 L1 규범을 채택합니다. 단일 생성기를 두 번 사용합니다. 먼저 원본 이미지를 대상 도메인의 이미지로 번역한 다음 번역된 이미지에서 원본 이미지를 재구성합니다.

Full Objective. Finally, the objective functions to optimize G and D are written, respectively, as전체 목표. 마지막으로 G와 D를 최적화하는 목적 함수는 각각 다음과 같이 작성됩니다.



where are hyper-parameters that control the relative importance of domain classification and reconstruction losses, respectively, compared to the adversarial loss. We use in all of our experiments.여기서 및 는 적대적 손실과 비교하여 도메인 분류 및 재구성 손실의 상대적 중요성을 각각 제어하는 ​​하이퍼 매개변수입니다. 우리는 모든 실험에서 및 을 사용합니다.

**3.2. Training with Multiple Datasets**

An important advantage of StarGAN is that it simultaneously incorporates multiple datasets containing different types of labels, so that StarGAN can control all the labels at the test phase. An issue when learning from multiple datasets, however, is that the label information is only partially known to each dataset. In the case of CelebA [19] and RaFD [13], while the former contains labels for attributes such as hair color and gender, it does not have any labels for facial expressions such as ‘happy’ and ‘angry’, and vice versa for the latter. This is problematic because the complete information on the label vector c’ is required when reconstructing the input image x from the translated image G(x, c) (See Eq. (4)).

StarGAN의 중요한 장점은 다른 유형의 레이블을 포함하는 여러 데이터 세트를 동시에 통합하므로 StarGAN이 테스트 단계에서 모든 레이블을 제어할 수 있다는 것입니다. 그러나 여러 데이터 세트에서 학습할 때의 문제는 레이블 정보가 각 데이터 세트에 부분적으로만 알려져 있다는 것입니다. CelebA[19]와 RaFD[13]의 경우 전자는 머리색, 성별 등의 속성에 대한 레이블이 포함되어 있지만 '행복', '화난'과 같은 표정 레이블은 없으며 그 반대의 경우도 마찬가지입니다. 후자를 위해. 이것은 번역된 이미지 G(x, c)에서 입력 이미지 x를 재구성할 때 레이블 벡터 c'에 대한 완전한 정보가 필요하기 때문에 문제가 됩니다(식 (4) 참조).

Mask Vector. To alleviate this problem, we introduce a mask vector m that allows StarGAN to ignore unspecified labels and focus on the explicitly known label provided by a particular dataset. In StarGAN, we use an n-dimensional one-hot vector to represent m, with n being the number of datasets. In addition, we define a unified version of the label as a vector

마스크 벡터입니다. 이 문제를 완화하기 위해 StarGAN이 지정되지 않은 레이블을 무시하고 특정 데이터 세트에서 제공하는 명시적으로 알려진 레이블에 집중할 수 있도록 하는 마스크 벡터 m을 도입합니다. StarGAN에서는 n차원 1-hot 벡터를 사용하여 m을 나타내며 n은 데이터 세트 수입니다. 또한 레이블의 통합 버전을 벡터로 정의합니다.



where [·] refers to concatenation, and represents a vector for the labels of the dataset. The vector of the known label ci can be represented as either a binary vector for binary attributes or a one-hot vector for categorical attributes. For the remaining unknown labels we simply assign zero values. In our experiments, we utilize the CelebA and RaFD datasets, where n is two.

여기서 [·]는 연결을 나타내고 는 i번째 데이터 세트의 레이블에 대한 벡터를 나타냅니다. 알려진 레이블 의 벡터는 이진 속성에 대한 이진 벡터 또는 범주 속성에 대한 원-핫 벡터로 나타낼 수 있습니다. 나머지 n-1개의 알려지지 않은 레이블에 대해서는 단순히 0 값을 할당합니다. 실험에서 CelebA 및 RaFD 데이터 세트를 활용합니다. 여기서 n은 2입니다.

Training Strategy. When training StarGAN with multiple datasets, we use the domain label defined in Eq. (7) as input to the generator. By doing so, the generator learns to ignore the unspecified labels, which are zero vectors, and focus on the explicitly given label. The structure of the generator is exactly the same as in training with a single dataset, except for the dimension of the input label . On the other hand, we extend the auxiliary classifier of the discriminator to generate probability distributions over labels for all datasets. Then, we train the model in a multi-task learning setting, where the discriminator tries to minimize only the classification error associated to the known label. For example, when training with images in CelebA, the discriminator minimizes only classification errors for labels related to CelebA attributes, and not facial expressions related to RaFD. Under these settings, by alternating between CelebA and RaFD the discriminator learns all of the discriminative features for both datasets, and the generator learns to control all the labels in both datasets.

훈련 전략. 여러 데이터 세트로 StarGAN을 훈련할 때 Eq.에 정의된 도메인 레이블 를 사용합니다. (7) 발전기에 대한 입력으로. 그렇게 함으로써 생성기는 0 벡터인 지정되지 않은 레이블을 무시하고 명시적으로 지정된 레이블에 초점을 맞추는 방법을 학습합니다. 생성기의 구조는 입력 레이블 의 차원을 제외하고 단일 데이터 세트를 사용한 훈련과 정확히 동일합니다. 다른 한편으로, 모든 데이터 세트에 대한 레이블에 대한 확률 분포를 생성하기 위해 판별자의 보조 분류기를 확장합니다. 그런 다음 판별자가 알려진 레이블과 관련된 분류 오류만 최소화하려고 시도하는 다중 작업 학습 설정에서 모델을 훈련합니다. 예를 들어 CelebA에서 이미지로 훈련할 때 판별자는 RaFD와 관련된 얼굴 표정이 아닌 CelebA 속성과 관련된 레이블에 대한 분류 오류만 최소화합니다. 이러한 설정에서 판별자는 CelebA와 RaFD를 번갈아 가며 두 데이터 세트에 대한 모든 판별 기능을 학습하고 생성자는 두 데이터 세트의 모든 레이블을 제어하는 ​​방법을 학습합니다.

텍스트, 가장, 그룹, 사람이(가) 표시된 사진

자동 생성된 설명

Figure 4. Facial attribute transfer results on the CelebA dataset. The first column shows the input image, next four columns show the single ttribute transfer results, and rightmost columns show the multi-attribute transfer results. H: Hair color, G: Gender, A: Aged.그림 4. CelebA 데이터 세트의 얼굴 속성 전송 결과. 첫 번째 열은 입력 이미지를 보여주고, 다음 4개의 열은 단일 속성 전송 결과를 보여주고, 가장 오른쪽 열은 다중 속성 전송 결과를 보여줍니다. H: 머리 색깔, G: 성별, A: 노화.

**4. Implementation**

Improved GAN Training. To stabilize the training process and generate higher quality images, we replace Eq. (1) with Wasserstein GAN objective with gradient penalty [1, 4] defined as

향상된 GAN 훈련. 훈련 과정을 안정화하고 더 높은 품질의 이미지를 생성하기 위해 Eq. (1) 기울기 페널티가 있는 Wasserstein GAN 대물렌즈 [1, 4]가 다음과 같이 정의됨

텍스트이(가) 표시된 사진

자동 생성된 설명

where is sampled uniformly along a straight line between a pair of a real and a generated images. We use for all experiments.여기서 는 한 쌍의 실제 이미지와 생성된 이미지 사이의 직선을 따라 균일하게 샘플링됩니다. 모든 실험에 을 사용합니다.

Network Architecture. Adapted from CycleGAN [33], StarGAN has the generator network composed of two convolutional layers with the stride size of two for downsampling, six residual blocks [5], and two transposed convolutional layers with the stride size of two for upsampling. We use instance normalization [29] for the generator but no normalization for the discriminator. We leverage PatchGANs [7, 15, 33] for the discriminator network, which classifies whether local image patches are real or fake. See the appendix (Section 7.2) for more details about the network architecture

네트워크 아키텍처. CycleGAN[33]에서 채택된 StarGAN은 다운샘플링을 위한 보폭 크기가 2인 2개의 컨볼루션 레이어, 6개의 잔여 블록, 업샘플링을 위한 보폭 크기가 2인 두 개의 전치된 컨볼루션 레이어로 구성된 생성기 네트워크를 가지고 있습니다. 생성기에는 인스턴스 정규화를 사용하지만 판별자에는 정규화를 사용하지 않습니다. 우리는 로컬 이미지 패치가 진짜인지 가짜인지를 분류하는 판별자 네트워크에 PatchGAN을 활용합니다. 네트워크 아키텍처에 대한 자세한 내용은 부록(섹션 7.2)을 참조하십시오.

**5. Experiments**

In this section, we first compare StarGAN against recent methods on facial attribute transfer by conducting user studies. Next, we perform a classification experiment on facial expression synthesis. Lastly, we demonstrate empirical results that StarGAN can learn image-to-image translation from multiple datasets. All our experiments were conducted by using the model output from unseen images during the training phase.이 섹션에서는 먼저 StarGAN을 사용자 연구를 수행하여 얼굴 속성 전송에 대한 최근 방법과 비교합니다. 다음으로 표정 합성에 대한 분류 실험을 수행합니다. 마지막으로 StarGAN이 여러 데이터 세트에서 이미지 대 이미지 번역을 학습할 수 있다는 실증적 결과를 보여줍니다. 우리의 모든 실험은 훈련 단계에서 보이지 않는 이미지의 모델 출력을 사용하여 수행되었습니다.

**5.1. Baseline Models**

As our baseline models, we adopt DIAT [16] and CycleGAN [33], both of which performs image-to-image translation between two different domains. For comparison, we trained these models multiple times for every pair of two different domains. We also adopt IcGAN [23] as a baseline which can perform attribute transfer using a cGAN [22].

기본 모델로 DIAT[16] 및 CycleGAN을 채택했으며 둘 다 서로 다른 두 도메인 간에 이미지 대 이미지 변환을 수행합니다. 비교를 위해 두 개의 서로 다른 도메인의 모든 쌍에 대해 이러한 모델을 여러 번 훈련했습니다. 또한 cGAN을 사용하여 속성 전송을 수행할 수 있는 기준선으로 IcGAN을 채택합니다.

DIAT uses an adversarial loss to learn the mapping from , where x and y are face images in two different domains X and Y, respectively. This method has a regularization term on the mapping as to preserve identity features of the source image, where F is a feature extractor pretrained on a face recognition task.DIAT는 x ∈X에서 y ∈Y로의 매핑을 학습하기 위해 적대적 손실을 사용합니다. 여기서 x와 y는 각각 두 개의 다른 도메인 X와 Y의 얼굴 이미지입니다. 이 방법은 매핑에 대한 정규화 용어를 로 사용하여 원본 이미지의 식별 특징을 보존합니다. 여기서 F는 얼굴 인식 작업에 대해 사전 훈련된 특징 추출기입니다.

CycleGAN also uses an adversarial loss to learn the mapping between two different domains X and Y . This method regularizes the mapping via cycle consistency losses, .This method requires two generators and discriminators for each pair of two different domains.

CycleGAN은 또한 두 개의 다른 도메인 X와 Y 사이의 매핑을 학습하기 위해 적대적 손실을 사용합니다. 이 방법은 주기 일관성 손실인 및 을 통해 매핑을 정규화합니다. 이 방법에는 서로 다른 두 도메인의 각 쌍에 대해 두 개의 생성자와 식별자가 필요합니다.

IcGAN combines an encoder with a cGAN [22] model. cGAN learns the mapping G : {z, c} → x that generates an image x conditioned on both the latent vector z and the conditional vector c. In addition, IcGAN introduces an encoder to learn the inverse mappings of cGAN, . This allows IcGAN to synthesis images by only changing the conditional vector and preserving the latent vector.

IcGAN은 인코더와 cGAN[22] 모델을 결합합니다. cGAN은 잠재 벡터 z와 조건 벡터 c 모두에 대해 조건이 지정된 이미지 x를 생성하는 매핑 G : {z, c} → x를 학습합니다. 또한 IcGAN은 인코더를 도입하여 cGAN, 및 의 역 매핑을 학습합니다. 이를 통해 IcGAN은 조건 벡터만 변경하고 잠재 벡터를 보존하여 이미지를 합성할 수 있습니다.

텍스트, 가장, 그룹이(가) 표시된 사진

자동 생성된 설명

Figure 5. Facial expression synthesis results on the RaFD dataset.그림 5. RaFD 데이터 세트의 얼굴 표정 합성 결과.

**5.2. Datasets**

CelebA. The CelebFaces Attributes (CelebA) dataset [19] contains 202,599 face images of celebrities, each annotated with 40 binary attributes. We crop the initial 178×218 size images to 178×178, then resize them as 128×128. We randomly select 2,000 images as test set and use all remaining images for training data. We construct seven domains using the following attributes: hair color (black, blond, brown), gender (male/female), and age (young/old).

셀레바. CelebFaces Attributes(CelebA) 데이터 세트에는 각각 40개의 바이너리 속성으로 주석이 달린 유명인의 얼굴 이미지 202,599개가 포함되어 있습니다. 초기 178×218 크기 이미지를 178×178로 자른 다음 128×128로 크기를 조정합니다. 무작위로 2,000개의 이미지를 테스트 세트로 선택하고 나머지 이미지를 모두 훈련 데이터로 사용합니다. 머리색(검정, 금발, 갈색), 성별(남성/여성), 연령(젊은/노인) 속성을 사용하여 7개의 도메인을 구성합니다.

RaFD. The Radboud Faces Database (RaFD) consists of 4,824 images collected from 67 participants. Each participant makes eight facial expressions in three different gaze directions, which are captured from three different angles. We crop the images to 256 × 256, where the faces are centered, and then resize them to 128 × 128.RaFD. Radboud Faces Database(RaFD)는 67명의 참가자로부터 수집한 4,824개의 이미지로 구성됩니다. 각 참가자는 세 가지 다른 각도에서 캡처한 세 가지 시선 방향에서 여덟 가지 표정을 만듭니다. 얼굴이 중앙에 있는 256 × 256으로 이미지를 자른 다음 128 × 128로 크기를 조정합니다.

**5.3. Training**

All models are trained using Adam [11] with . For data augmentation we flip the images horizontally with a probability of 0.5. We perform one generator update after five discriminator updates as in [4]. The batch size is set to 16 for all experiments. For experiments on CelebA, we train all models with a learning rate of 0.0001 for the first 10 epochs and linearly decay the learning rate to 0 over the next 10 epochs. To compensate for the lack of data, when training with RaFD we train all models for 100 epochs with a learning rate of 0.0001 and apply the same decaying strategy over the next 100 epochs. Training takes about one day on a single NVIDIA Tesla M40 GPU.

모든 모델은 β\_1=0.5 및 β\_2=0.999인 Adam을 사용하여 훈련되었습니다. 데이터 증대를 위해 0.5의 확률로 이미지를 수평으로 뒤집습니다. [4]와 같이 5번의 판별자 업데이트 후에 1번의 생성자 업데이트를 수행합니다. 배치 크기는 모든 실험에 대해 16으로 설정됩니다. CelebA에 대한 실험을 위해 처음 10개 epoch 동안 0.0001의 학습률로 모든 모델을 훈련하고 다음 10개 epoch 동안 학습률을 0으로 선형 감쇠합니다. 데이터 부족을 보완하기 위해 RaFD로 훈련할 때 학습률이 0.0001인 100세대 동안 모든 모델을 교육하고 다음 100세대 동안 동일한 감쇠 전략을 적용합니다. 훈련은 단일 NVIDIA Tesla M40 GPU에서 하루 정도 걸립니다.

**5.4. Experimental Results on CelebA**

We first compare our proposed method to the baseline models on a single and multi-attribute transfer tasks. We train the cross-domain models such as DIAT and CycleGAN multiple times considering all possible attribute value pairs. In the case of DIAT and CycleGAN, we perform multi-step translations to synthesize multiple attributes (e.g. transferring a gender attribute after changing a hair color).먼저 제안된 방법을 단일 및 다중 속성 전송 작업에 대한 기준 모델과 비교합니다. 가능한 모든 속성 값 쌍을 고려하여 DIAT 및 CycleGAN과 같은 교차 도메인 모델을 여러 번 훈련합니다. DIAT 및 CycleGAN의 경우 여러 속성을 합성하기 위해 다단계 번역을 수행합니다(예: 헤어 컬러 변경 후 성별 속성 전송).

Qualitative evaluation. Fig. 4 shows the facial attribute transfer results on CelebA. We observed that our method provides a higher visual quality of translation results on test data compared to the cross-domain models. One possible reason is the regularization effect of StarGAN through a multi-task learning framework. In other words, rather than training a model to perform a fixed translation (e.g., brownto-blond hair), which is prone to overfitting, we train our model to flexibly translate images according to the labels of the target domain. This allows our model to learn reliable features universally applicable to multiple domains of images with different facial attribute values.

정성적 평가. 그림 4는 CelebA의 얼굴 속성 전달 결과를 보여준다. 우리는 우리의 방법이 교차 도메인 모델에 비해 테스트 데이터에 대한 번역 결과의 시각적 품질이 더 높다는 것을 관찰했습니다. 한 가지 가능한 이유는 다중 작업 학습 프레임워크를 통한 StarGAN의 정규화 효과입니다. 즉, 과적합되기 쉬운 고정 번역(예: 갈색에서 금발)을 수행하도록 모델을 훈련하는 대신 대상 도메인의 레이블에 따라 유연하게 이미지를 번역하도록 모델을 훈련합니다. 이를 통해 우리 모델은 얼굴 속성 값이 다른 여러 이미지 도메인에 보편적으로 적용할 수 있는 안정적인 기능을 학습할 수 있습니다.

Furthermore, compared to IcGAN, our model demonstrates an advantage in preserving the facial identity feature of an input. We conjecture that this is because our method maintains the spatial information by using activation maps from the convolutional layer as latent representation, rather than just a low-dimensional latent vector as in IcGAN.또한 IcGAN과 비교하여 우리 모델은 입력의 얼굴 식별 기능을 보존하는 이점을 보여줍니다. 우리는 이것이 우리의 방법이 IcGAN에서와 같은 저차원 잠재 벡터가 아니라 잠재 표현으로 컨볼루션 레이어의 활성화 맵을 사용하여 공간 정보를 유지하기 때문이라고 추측합니다.

Quantitative evaluation protocol. For quantitative evaluations, we performed two user studies in a survey format using Amazon Mechanical Turk (AMT) to assess single and multiple attribute transfer tasks. Given an input image, the Turkers were instructed to choose the best generated image based on perceptual realism, quality of transfer in attribute(s), and preservation of a figure’s original identity. The options were four randomly shuffled images generated from four different methods. The generated images in one study have a single attribute transfer in either hair color (black, blond, brown), gender, or age. In another study, the generated images involve a combination of attribute transfers. Each Turker was asked 30 to 40 questions with a few simple yet logical questions for validating human effort. The number of validated Turkers in each user study is 146 and 100 in single and multiple transfer tasks, respectively.

정량적 평가 프로토콜. 정량적 평가를 위해 Amazon Mechanical Turk(AMT)를 사용하여 단일 및 다중 속성 전송 작업을 평가하는 설문조사 형식으로 두 가지 사용자 연구를 수행했습니다. 입력 이미지가 주어지면 Turkers는 지각적 사실주의, 속성 전달의 품질 및 인물의 원래 정체성 보존을 기반으로 가장 잘 생성된 이미지를 선택하도록 지시받았습니다. 옵션은 네 가지 다른 방법에서 생성된 네 개의 무작위로 섞인 이미지였습니다. 한 연구에서 생성된 이미지에는 머리색(검은색, 금발, 갈색), 성별 또는 연령에 대한 단일 속성 전송이 있습니다. 다른 연구에서 생성된 이미지는 속성 전송의 조합을 포함합니다. 각 Turker는 인간의 노력을 검증하기 위한 몇 가지 간단하면서도 논리적인 질문과 함께 30~40개의 질문을 받았습니다. 각 사용자 연구에서 검증된 Turkers의 수는 단일 및 다중 전송 작업에서 각각 146 및 100입니다.

텍스트, 사람, 가장, 그룹이(가) 표시된 사진

자동 생성된 설명

Figure 6. Facial expression synthesis results of StarGAN-SNG and StarGAN-JNT on CelebA dataset.

그림 6. CelebA 데이터 세트에서 StarGAN-SNG 및 StarGAN-JNT의 얼굴 표정 합성 결과.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 1. AMT perceptual evaluation for ranking different models on a single attribute transfer task. Each column sums to 100%.표 1. 단일 속성 전송 작업에서 서로 다른 모델의 순위를 매기기 위한 AMT 지각 평가. 각 열의 합계는 100%입니다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 2. AMT perceptual evaluation for ranking different models on a multi-attribute transfer task. H: Hair color; G: Gender; A: Aged.표 2. 다중 속성 전송 작업에서 서로 다른 모델의 순위를 매기기 위한 AMT 지각 평가. H: 머리 색깔; G: 성별; A: 늙었다.

Quantitative results. Tables 1 and 2 show the results of our AMT experiment on single- and multi-attribute transfer tasks, respectively. StarGAN obtained the majority of votes for best transferring attributes in all cases. In the case of gender changes in Table 1, the voting difference between our model and other models was marginal, e.g., 39.1% for StarGAN vs. 31.4% for DIAT. However, in multi-attribute changes, e.g., the ‘G+A’ case in Table 2, the performance difference becomes significant, e.g., 49.8% for StarGAN vs. 20.3% for IcGAN), clearly showing the advantages of StarGAN in more complicated, multi-attribute transfer tasks. This is because unlike the other methods, StarGAN can handle image translation involving multiple attribute changes by randomly generating a target domain label in the training phase.정량적 결과. 표 1과 2는 각각 단일 및 다중 속성 전송 작업에 대한 AMT 실험의 결과를 보여줍니다. StarGAN은 모든 경우에 최고의 속성 전송에 대해 과반수를 얻었습니다. 표 1의 성별 변경의 경우 우리 모델과 다른 모델 간의 투표 차이는 미미했습니다(예: StarGAN의 경우 39.1% 대 DIAT의 경우 31.4%). 그러나 다중 속성 변경(예: 표 2의 'G+A' 경우)에서는 성능 차이가 크게 나타나(예: StarGAN의 경우 49.8% vs. IcGAN의 경우 20.3%) 더 복잡한 시스템에서 StarGAN의 장점을 명확하게 보여줍니다. 다중 속성 전송 작업. StarGAN은 다른 방법과 달리 훈련 단계에서 대상 도메인 레이블을 무작위로 생성하여 여러 속성 변경이 포함된 이미지 번역을 처리할 수 있기 때문입니다.

**5.5. Experimental Results on RaFD**

We next train our model on the RaFD dataset to learn the

task of synthesizing facial expressions. To compare StarGAN and baseline models, we fix the input domain as the ‘neutral’ expression, but the target domain varies among the seven remaining expressions.

표정 합성 작업. StarGAN과 베이스라인 모델을 비교하기 위해 입력 영역을 '중립' 표현으로 고정하지만 대상 영역은 나머지 7개 표현 사이에서 다릅니다.

Qualitative evaluation. As seen in Fig. 5, StarGAN clearly generates the most natural-looking expressions while properly maintaining the personal identity and facial features of the input. While DIAT and CycleGAN mostly preserve the identity of the input, many of their results are shown blurry and do not maintain the degree of sharpness as seen in the input. IcGAN even fails to preserve the personal identity in the image by generating male images.

정성적 평가. 그림 5에서 볼 수 있듯이 StarGAN은 입력의 개인 신원과 얼굴 특징을 적절하게 유지하면서 가장 자연스러운 표정을 명확하게 생성합니다. DIAT 및 CycleGAN은 대부분 입력의 ID를 유지하지만 많은 결과가 흐릿하게 표시되고 입력에서 볼 수 있는 선명도를 유지하지 않습니다. IcGAN은 심지어 남성 이미지를 생성하여 이미지에서 개인의 정체성을 유지하는 데 실패합니다.

We believe that the superiority of StarGAN in the image quality is due to its implicit data augmentation effect from a multi-task learning setting. RaFD images contain a relatively small size of samples, e.g., 500 images per domain. When trained on two domains, DIAT and CycleGAN can only use 1,000 training images at a time, but StarGAN can use 4,000 images in total from all the available domains for its training. This allows StarGAN to properly learn how to maintain the quality and sharpness of the generated output.

이미지 품질에서 StarGAN의 우수성은 다중 작업 학습 설정에서 암시적인 데이터 증대 효과 때문이라고 생각합니다. RaFD 이미지에는 도메인당 500개의 이미지와 같이 비교적 작은 크기의 샘플이 포함됩니다. 두 도메인에서 훈련할 때 DIAT와 CycleGAN은 한 번에 1,000개의 훈련 이미지만 사용할 수 있지만 StarGAN은 훈련에 사용 가능한 모든 도메인에서 총 4,000개의 이미지를 사용할 수 있습니다. 이를 통해 StarGAN은 생성된 출력의 품질과 선명도를 유지하는 방법을 적절하게 학습할 수 있습니다.

Quantitative evaluation. For a quantitative evaluation, we compute the classification error of a facial expression on synthesized images. We trained a facial expression classifier on the RaFD dataset (90%/10% splitting for training and test sets) using a ResNet-18 architecture [5], resulting in a near-perfect accuracy of 99.55%. We then trained each of image translation models using the same training set and performed image translation on the same, unseen test set. Finally, we classified the expression of these translated images using the above-mentioned classifier. As can be seen in Table 3, our model achieves the lowest classification error, indicating that our model produces the most realistic facial expressions among all the methods compared.

정량적 평가. 정량적 평가를 위해 합성 영상에서 표정의 분류 오류를 계산합니다. 우리는 ResNet-18 아키텍처를 사용하여 RaFD 데이터 세트(훈련 및 테스트 세트에 대해 90%/10% 분할)에 대해 얼굴 표정 분류기를 훈련하여 99.55%의 거의 완벽한 정확도를 얻었습니다. 그런 다음 동일한 훈련 세트를 사용하여 각 이미지 번역 모델을 훈련하고 동일한 보이지 않는 테스트 세트에서 이미지 번역을 수행했습니다. 마지막으로 위에서 언급한 분류기를 사용하여 이러한 번역된 이미지의 표현을 분류했습니다. Table 3에서 볼 수 있듯이 우리 모델은 가장 낮은 분류 오차를 보이며 비교한 모든 방법 중 가장 사실적인 얼굴 표정을 생성함을 알 수 있다.

테이블이(가) 표시된 사진

자동 생성된 설명

Table 3. Classification errors [%] and the number of parameters on the RaFD dataset.

표 3. 분류 오류[%] 및 RaFD 데이터 세트의 매개변수 수.

Another important advantage of our model is the scalability in terms of the number of parameters required. The last column in Table 3 shows that the number of parameters required to learn all translations by StarGAN is seven times smaller than that of DIAT and fourteen times smaller than that of CycleGAN. This is because StarGAN requires only a single generator and discriminator pair, regardless of the number of domains, while in the case of cross-domain models such as CycleGAN, a completely different model should be trained for each source-target domain pair.

우리 모델의 또 다른 중요한 이점은 필요한 매개변수의 수 측면에서 확장성입니다. 표 3의 마지막 열은 StarGAN에 의한 모든 번역을 학습하는 데 필요한 매개변수의 수가 DIAT보다 7배, CycleGAN보다 14배 작음을 보여줍니다. StarGAN은 도메인 수에 관계없이 하나의 생성자와 판별자 쌍만 필요하지만 CycleGAN과 같은 교차 도메인 모델의 경우 각 소스-타겟 도메인 쌍에 대해 완전히 다른 모델을 학습해야 하기 때문입니다.

**5.6. Experimental Results on CelebA+RaFD**

Finally, we empirically demonstrate that our model can learn not only from multiple domains within a single dataset, but also from multiple datasets. We train our model jointly on the CelebA and RaFD datasets using the mask vector (see Section 3.2). To distinguish between the model trained only on RaFD and the model trained on both CelebA and RaFD, we denote the former as StarGAN-SNG (single) and the latter as StarGAN-JNT (joint).

마지막으로, 우리는 우리 모델이 단일 데이터 세트 내의 여러 도메인뿐만 아니라 여러 데이터 세트에서도 학습할 수 있음을 경험적으로 보여줍니다. 우리는 마스크 벡터를 사용하여 CelebA 및 RaFD 데이터 세트에 대해 모델을 공동으로 훈련합니다(섹션 3.2 참조). RaFD에서만 훈련된 모델과 CelebA와 RaFD 모두에서 훈련된 모델을 구별하기 위해 전자를 StarGAN-SNG(단일)로, 후자를 StarGAN-JNT(조인트)로 표시합니다.

Effects of joint training. Fig. 6 shows qualitative comparisons between StarGAN-SNG and StarGAN-JNT, where the task is to synthesize facial expressions of images in CelebA. StarGAN-JNT exhibits emotional expressions with high visual quality, while StarGAN-SNG generates reasonable but blurry images with gray backgrounds. This difference is due to the fact that StarGAN-JNT learns to translate CelebA images during training but not StarGAN-SNG. In other words, StarGAN-JNT can leverage both datasets to improve shared low-level tasks such facial keypoint detection and segmentation. By utilizing both CelebA and RaFD, StarGAN-JNT can improve these low-level tasks, which is beneficial to learning facial expression synthesis.

합동 훈련의 효과. 그림 6은 StarGAN-SNG와 StarGAN-JNT 간의 정성적 비교를 보여주고 있으며, 여기서 과제는 CelebA에서 이미지의 얼굴 표정을 합성하는 것이다. StarGAN-JNT는 높은 화질로 감성적인 표현을, StarGAN-SNG는 회색 배경으로 합리적이면서도 흐릿한 이미지를 연출합니다. 이 차이는 StarGAN-JNT가 훈련 중에 CelebA 이미지를 번역하지만 StarGAN-SNG는 번역하지 않는다는 사실에 기인합니다. 즉, StarGAN-JNT는 두 데이터 세트를 모두 활용하여 안면 키포인트 감지 및 세분화와 같은 공유 하위 수준 작업을 개선할 수 있습니다. CelebA와 RaFD를 모두 활용하여 StarGAN-JNT는 이러한 낮은 수준의 작업을 개선할 수 있으며 이는 얼굴 표정 합성을 학습하는 데 도움이 됩니다.

텍스트, 가장, 사람, 그룹이(가) 표시된 사진

자동 생성된 설명

Figure 7. Learned role of the mask vector. All images are generated by StarGAN-JNT. The first row shows the result of applying the proper mask vector, and the last row shows the result of applying the wrong mask vector.

그림 7. 마스크 벡터의 학습된 역할. 모든 이미지는 StarGAN-JNT에 의해 생성됩니다. 첫 번째 행은 적절한 마스크 벡터를 적용한 결과이고 마지막 행은 잘못된 마스크 벡터를 적용한 결과입니다.

Learned role of mask vector. In this experiment, we gave a one-hot vector c by setting the dimension of a particular facial expression (available from the second dataset, RaFD) to one. In this case, since the label associated with the second data set is explicitly given, the proper mask vector would be [0, 1]. Fig. 7 shows the case where this proper mask vector was given and the opposite case where a wrong mask vector of [1, 0] was given. When the wrong mask vector was used, StarGAN-JNT fails to synthesize facial expressions, and it manipulates the age of the input image. This is because the model ignores the facial expression label as unknown and treats the facial attribute label as valid by the mask vector. Note that since one of the facial attributes is ‘young’, the model translates the image from young to old when it takes in a zero vector as input. From this behavior, we can confirm that StarGAN properly learned the intended role of a mask vector in image-to-image translations when involving all the labels from multiple datasets altogether.

마스크 벡터의 역할을 배웠습니다. 이 실험에서는 특정 얼굴 표정의 차원(두 번째 데이터 세트 RaFD에서 사용 가능)을 1로 설정하여 원-핫 벡터 c를 제공했습니다. 이 경우 두 번째 데이터 세트와 관련된 레이블이 명시적으로 제공되므로 적절한 마스크 벡터는 [0, 1]이 됩니다. 그림 7은 이 적절한 마스크 벡터가 주어진 경우와 [1, 0]의 잘못된 마스크 벡터가 주어진 반대의 경우를 보여줍니다. 잘못된 마스크 벡터를 사용하면 StarGAN-JNT가 얼굴 표정 합성에 실패하고 입력 이미지의 나이를 조작합니다. 이는 모델이 얼굴 표정 레이블을 알 수 없는 것으로 무시하고 얼굴 속성 레이블을 마스크 벡터에 의해 유효한 것으로 취급하기 때문입니다. 얼굴 속성 중 하나가 '젊음'이기 때문에 모델은 0 벡터를 입력으로 받을 때 이미지를 젊음에서 노인으로 변환합니다. 이 동작에서 StarGAN이 여러 데이터 세트의 모든 레이블을 모두 포함할 때 이미지-이미지 변환에서 마스크 벡터의 의도된 역할을 적절하게 학습했음을 확인할 수 있습니다.

**6. Conclusion**

In this paper, we proposed StarGAN, a scalable imageto-image translation model among multiple domains using a single generator and a discriminator. Besides the advantages in scalability, StarGAN generated images of higher visual quality compared to existing methods [16, 23, 33], owing to the generalization capability behind the multi-task learning setting. In addition, the use of the proposed simple mask vector enables StarGAN to utilize multiple datasets with different sets of domain labels, thus handling all available labels from them. We hope our work to enable users to develop interesting image translation applications across multiple domains.

본 논문에서는 단일 생성기와 판별기를 사용하여 여러 도메인 간의 확장 가능한 이미지 대 이미지 변환 모델인 StarGAN을 제안했습니다. StarGAN은 확장성의 장점 외에도 다중 작업 학습 설정의 일반화 기능으로 인해 기존 방법[16, 23, 33]에 비해 더 높은 시각적 품질의 이미지를 생성했습니다. 또한 제안된 단순 마스크 벡터를 사용하면 StarGAN이 서로 다른 도메인 레이블 세트를 가진 여러 데이터 세트를 활용하여 사용 가능한 모든 레이블을 처리할 수 있습니다. 사용자가 여러 도메인에서 흥미로운 이미지 번역 응용 프로그램을 개발할 수 있도록 우리의 작업을 바랍니다.

Acknowledgements. This work was mainly done while the first author did a research internship at Clova AI Research, NAVER. We thank all the researchers at NAVER, especially Donghyun Kwak, for insightful discussions. This work was partially supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIP) (No. NRF2016R1C1B2015924). Jaegul Choo is the corresponding author.감사합니다. 이 작업은 제1저자가 네이버 클로바 AI 리서치에서 리서치 인턴쉽을 하면서 주로 이루어졌습니다. 깊이 있는 토론을 해주신 네이버 연구원들, 특히 곽동현 분들께 감사드립니다. 이 작업은 한국정부(MSIP)의 지원을 받는 한국연구재단(NRF) 연구비(No. NRF2016R1C1B2015924)의 일부 지원을 받았습니다. 추재걸 교신저자.