Image Manipulation with Perceptual Discriminators

Diana Sungatullina*, Egor Zakharov*, Dmitry Ulyanov, and Victor Lempitsky

Skolkovo Institute of Science and Technology, Moscow, Russia
{d.sungatullina, egor.zakharov, dmitry.ulyanov, lempitsky}@skoltech.ru

Abstract. Systems that perform image manipulation using deep convolutional networks have achieved remarkable realism. Perceptual losses and losses based on adversarial discriminators are the two main classes of learning objectives behind these advances. In this work, we show how these two ideas can be combined in a principled and non-additive manner for unaligned image translation tasks. This is accomplished through a special architecture of the discriminator network inside generative adversarial learning framework. The new architecture, that we call a *perceptual discriminator*, embeds the convolutional parts of a pre-trained deep classification network inside the discriminator network. The resulting architecture can be trained on unaligned image datasets, while benefiting from the robustness and efficiency of perceptual losses. We demonstrate the merits of the new architecture in a series of qualitative and quantitative comparisons with baseline approaches and state-of-the-art frameworks for unaligned image translation.

Keywords: Image translation · Image editing · Perceptual loss · Generative adversarial networks

1 Introduction

Generative convolutional neural networks have achieved remarkable success in image manipulation tasks both due to their ability to train on large amount of data [20,23,12] and due to natural image priors associated with such architectures [38]. Recently, the ability to train image manipulation ConvNets has been shown in the *unaligned* training scenario [42,43,5], where the training is based on sets of images annotated with the presence/absence of a certain attribute, rather than based on *aligned* datasets containing {input,output} image pairs. The ability to train from unaligned data provides considerable flexibility in dataset collection and in learning new manipulation effects, yet poses additional algorithmic challenges.

Generally, the realism of the deep image manipulation methods is known to depend strongly on the choice of the loss functions that are used to train generative ConvNets. In particular, simplistic pixelwise losses (e.g. the squared

* indicates equal contribution

distance loss) are known to limit the realism and are also non-trivial to apply in the unaligned training scenario. The rapid improvement of realism of deep image generation and processing is thus associated with two classes of loss functions that go beyond pixel-wise losses. The first group (so-called *perceptual losses*) is based on matching activations inside pre-trained deep convolutional networks (the VGG architecture trained for ILSVRC image classification is by far the most popular choice [35]). The second group consists of *adversarial losses*, where the loss function is defined implicitly using a separate *discriminator* network that is trained adversarially in parallel with the main generative network.

The two groups (perceptual losses and adversarial losses) are known to have largely complementary strengths and weaknesses. Thus, perceptual losses are easy to incorporate and are easy to scale to high-resolution images; however, their use in unaligned training scenario is difficult, as these loss terms require a concrete target image to match the activations to. Adversarial losses have the potential to achieve higher realism and can be used naturally in the unaligned scenarios, yet adversarial training is known to be hard to set up properly, often suffers from mode collapse, and is hard to scale to high-resolution images. Combining perceptual and adversarial losses in an additive way has been popular [11,40,24,33]. Thus, a generative ConvNet can be trained by minimizing a linear combination of an adversarial and a perceptual (and potentially some other) losses. Yet such additive combination includes not only strengths but also weaknesses of the two approaches. In particular, the use of a perceptual loss still incurs the use of aligned datasets for training.

In this work we present an architecture for realistic image manipulation, which combines perceptual and adversarial losses in a natural *non-additive* way. Importantly, the architecture keeps the ability of adversarial losses to train on unaligned datasets, while also benefits from the stability of perceptual losses. Our idea is very simple and concerned with the particular design of the discriminator network for adversarial training. The design encapsulates a pretrained classification network as the initial part of the discriminator. During adversarial training, the generator network is effectively learned to match the activations inside several layers of this reference network, just like the perceptual losses do. We show that the incorporation of the pretrained network into the discriminator stabilizes the training and scales well to higher resolution images, as is common with perceptual losses. At the same time, the use of adversarial training allows to avoid the need for aligned training data.

Generally, we have found that the suggested architecture can be trained with little tuning to impose complex image manipulations, such as adding to and removing smile from human faces, face ageing and rejuvenation, gender change, hair style change, etc. In the experiments, we show that our architecture can be used to perform complex manipulations at medium and high resolutions, and compare the proposed architecture with several adversarial learning-based baselines and recent methods for learning-based image manipulation.

2 Related work

Generative ConvNets. Our approach is related to a rapidly growing body of works on ConvNets for image generation and editing. Some of the earlier important papers on ConvNet image generation [12] and image processing [20,10,23] used per-pixel loss functions and fully supervised setting, so that at test time the target image is known for each input. While this demonstrated the capability of ConvNets to generate realistic images, the proposed systems all had to be trained on aligned datasets and the amount of high-frequency details in the output images was limited due to deficiencies of pixel-wise loss functions.

Perceptual Losses. The work of Mahendran and Vedaldi [28] has demonstrated that the activations invoked by an image within a pre-trained convolutional network can be used to recover the original image. Gatys et al. [13] showed that such activations can serve as content descriptors or texture descriptors of the input image, while Dosovitsky and Brox [11], Ulyanov et al. [37], Johnson et al. [21] have shown that the mismatches between the produced and the target activations can be used as so-called *perceptual losses* for a generative ConvNet. The recent work of [7] pushed the spatial resolution and the realism of images produced by a feed-forward ConvNet with perceptual losses to megapixel resolution. Generally, in all the above-mentioned works [7,37,21,11], the perceptual loss is applied in a fully supervised manner as for each training example the specific target deep activations (or the Gram matrix thereof) are given explicitly. Finally, [39] proposed a method that manipulates carefully aligned face images at high resolution by compositing desired activations of a deep pretrained network and finding an image that matches such activations using the non-feedforward optimization process similar to [28,13].

Adversarial Training. The most impressive results of generative ConvNets were obtained within generative adversarial networks (GANs) framework proposed originally by Goodfellow et al. [14]. The idea of adversarial training is to implement the loss function as a separate trainable network (the *discriminator*), which is trained in parallel and in adversarial way with the generative ConvNet (the *generator*). Multiple follow-up works including [30,34,3,22] investigated the choice of convolutional architectures for the generator and for the discriminator. Achieving reliable and robust convergence of generator-discriminator pairs remains challenging [15,8,27], and in particular requires considerably more efforts than training with perceptual loss functions.

Unaligned Adversarial Training. While a lot of the original interest to GANs was associated with unconditional image generation, recently the emphasis has shifted to the conditional image synthesis. Most relevant to our work are adversarially-trained networks that perform image translation, i.e. generate output images conditioned on input images. While initial methods used aligned

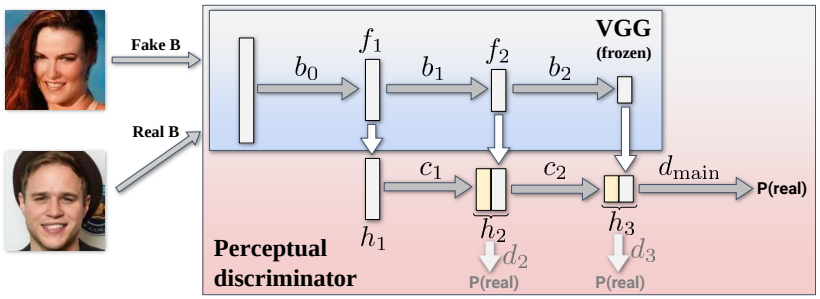

Fig. 1: The perceptual discriminator is composed of a pre-trained image classification network (such as VGG), split into blocks b_i. The parameters of those blocks are not changed during training, thus the discriminator retains access to so-called perceptual features. The outputs of these blocks are processed using learnable blocks of convolutional operations c_i and the outputs of those are used to predict the probability of an image being real or manipulated (the simpler version uses a single discriminator d_{main}, while additional path discriminators are used in the full version).

datasets for training [41,19], recently some impressive results have been obtained using unaligned training data, where only empirical distributions of the input and the output images are provided [42,5,43]. For face image manipulation, systems using adversarial training on unaligned data have been proposed in [6,9]. While we also make an emphasis on face manipulation, our contribution is orthogonal to [6,9] as perceptual discriminators can be introduced into their systems.

Combining Perceptual and Adversarial Losses. A growing number of works [11,24,40] use the combination of perceptual and adversarial loss functions to accomplish more stable training and to achieve convincing image manipulation at high resolution. Most recently, [33] showed that augmenting perceptual loss with the adversarial loss improves over the baseline system [7] (that has already achieved very impressive results) in the task of megapixel-sized conditional image synthesis. Invariably, the combination of perceptual and adversarial losses is performed in an additive manner, i.e. the two loss functions are weighted and added to each other (and potentially to some other terms). While such additive combination is simple and often very efficient, it limits learning to the aligned scenario, as perceptual terms still require to specify target activations for each training example. In this work, we propose a natural non-additive combination of perceptual losses and adversarial training that avoids the need for aligned data during training.

3 Perceptual discriminators

3.1 Background and motivation

Generative adversarial networks have shown impressive results in photorealistic image synthesis. The model includes a generative network G, that is trained to match the target distribution $p_{\text{target}}(\mathbf{y})$ in the data space $\mathcal{Y}$, and a discriminator network D that is trained to distinguish whether the input is real or generated by G. In the simplest form, the two networks optimize the policy function $V(D,G)$:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{y}\sim p_{\text{target}}(\mathbf{y})} \log D(\mathbf{y}) + \mathbb{E}_{\mathbf{x}\sim p_{\text{source}}(\mathbf{x})}[\log(1 - D(G(\mathbf{x}))], \quad (1)$$

In (1), the source distribution $p_{\text{source}}(\mathbf{x})$ may correspond to a simple parametric distribution in a latent space such as the unit Gaussian, so that after training unconditional samples from the learned approximation to $p_{\text{target}}(\mathbf{y})$ can be drawn. Alternatively, $p_{\text{source}}(\mathbf{x})$ may correspond to another empirical distribution in the image space $\mathcal{X}$. In this case, the generator learns to *translate* images from $\mathcal{X}$ to $\mathcal{Y}$, or to *manipulate* images in the space $\mathcal{X}$ (when it coincides with $\mathcal{Y}$). Although our contribution (perceptual discriminators) is applicable to both unconditional synthesis and image manipulation/translation, we focus our evaluation on the latter scenario. For the low resolution datasets, we use the standard non-saturating GAN modification, where the generator maximizes the log-likelihood of the discriminator instead of minimizing the objective (1) [14]. For high-resolution images, following CycleGAN [42], we use the LSGAN formulation [29].

Converging to good equilibria for any of the proposed GAN games is known to be hard [15,8,27]. In general, the performance of the trained generator network crucially depends on the architecture of the discriminator network, that needs to learn meaningful statistics, which are good for matching the target distribution p_{target}. The typical failure mode of GAN training is when the discriminator does not manage to learn such statistics before being "overpowered" by the generator.

3.2 Perceptual Discriminator Architecture

Multiple approaches have suggested to use activations invoked by an image $\mathbf{y}$ inside a deep pre-trained classification network $F(\mathbf{y})$ as statistics for such tasks as retrieval [4] or few-shot classification [31]. Mahendran and Vedaldi [28] have shown that activations computed after the convolutional part of such network retain most of the information about the input $\mathbf{y}$, i.e. are essentially invertable. Subsequent works such as [13,37,21,11] all used such "perceptual" statistics to match low-level details such as texture content, certain image resolution, or particular artistic style.

Following this line of work, we suggest to base the GAN discriminator $D(\mathbf{y})$ on the perceptual statistics computed by the reference network F on the input image $\mathbf{y}$, which can be either real (coming from p_{target}) or fake (produced by the generator). Our motivation is that a discriminator that uses perceptual features

has a better chance to learn good statistics than a discriminator initialized to a random network. For simplicity, we assume that the network F has a chain structure, e.g. F can be the VGGNet of [35].

Consider the subsequent blocks of the convolutional part of the reference network F, and denote them as $b_0, b_1, \dots, b_{K-1}$. Each block may include one or more convolutional layers interleaved with non-linearities and pooling operations. Then, the perceptual statistics $\{f_1(\mathbf{y}), \dots, f_K(\mathbf{y})\}$ are computed as:

$$f_1(\mathbf{y}) = b_0(\mathbf{y}) \tag{2}$$

$$f_i(\mathbf{y}) = b_{i-1}(f_{i-1}(\mathbf{y})), \quad i = 2, \dots, K\,, \tag{3}$$

so that each $f_i(\mathbf{y})$ is a stack of convolutional maps of the spatial dimensions $W_i \times W_i$. The dimension W_i is determined by the preceeding size W_{i-1} as well as by the presence of strides and pooling operations inside b_i. In our experiments we use features from consecutive blocks, i.e. $W_i = W_{i-1}/2$.

The overall structure of our discriminator is shown in Figure 1. The key novelty of our discriminator is the in-built perceptual statistics f_i (top of the image), which are known to be good at assessing image realism [13,21,39]. During the backpropagation, the gradients to the generator flow through the perceptual statistics extractors b_i, but the parameters of b_i are frozen and inherited from the network pretrained for large-scale classification. This stabilizes the training, and ensures that at each moment of time the discriminator has access to "good" features, and therefore cannot be overpowered by the generator easily.

In more detail, the proposed discriminator architecture combines together perceptual statistics using the following computations:

$$h_1(\mathbf{y}) = f_1(\mathbf{y}) \tag{4}$$

$$h_i(\mathbf{y}) = \texttt{stack}\left[c_{i-1}(h_{i-1}(\mathbf{y}), \phi_{i-1})\,,\, f_i(\mathbf{y})\right], \quad i = 2, \dots, K\,, \tag{5}$$

where `stack` denotes stacking operation, and the convolutional blocks c_j with learnable parameters ϕ_j (for $j = 1, \dots, K-1$) are composed of convolutions, leaky ReLU nonlinearities, and average pooling operations. Each of the c_j blocks thus transforms map stacks of the spatial size $W_j \times W_j$ to map stacks of the spatial size $W_{j+1} \times W_{j+1}$. Thus, the strides and pooling operations inside c_j match the strides and/or pooling operations inside b_j.

Using a series of convolutional and fully-connected layers with learnable parameters ψ_{main} applied to the representation $h_K(\mathbf{y})$, the discriminator outputs the probability d_{main} of the whole image $\mathbf{y}$ being real. For low- to medium-resolution images we perform experiments using only this probability. For high-resolution, we found that additional outputs from the discriminator resulted in better outcomes. Using the "patch discriminator" idea [19,42], to several feature representations h_j we apply a convolution+LeakyReLU block d_j with learnable parameters ψ_j that outputs probabilities $d_{j,p}$ at every spatial locations p. We then replace the regular log probability $\log D(\mathbf{y}) \equiv \log d_{\text{main}}$ of an image being real with:

$$\log D(\mathbf{y}) = \log d_{\text{main}}(\mathbf{y}) + \sum_j \sum_{p \in \text{Grid}(W_j \times W_j)} \log d_{j,p}(\mathbf{y}) \tag{6}$$

Note, that this makes our discriminator "multi-scale", since spatial resolution W_j varies for different j. The idea of multiple classifiers inside the discriminator have also been proposed recently in [40,18]. Unlike [40,18] where these classifiers are disjoint, in our architecture all such classifiers are different branches of the same network that has perceptual features underneath.

During training, the parameters of the c blocks inside the feature network F remain fixed, while the parameters ϕ_i of feature extractors c_i and the parameters ψ_i of the discriminators d_i are updated during the adversarial learning, which forces the "perceptual" alignment between the output of the generator and p_{target}. Thus, wrapping perceptual loss terms into additional layers c_i and d_i and putting them together into the adversarial discriminator allows us to use such perceptual terms in the unaligned training scenario. Such unaligned training was, in general, not possible with the "traditional" perceptual losses.

3.3 Architecture Details

Reference Network. Following multiple previous works [13,37,21], we consider the so-called *VGG network* from [35] trained on ILSVRC2012 [32] as the reference network F. In particular, we pick the VGG-19 variant, to which we simply refer to as VGG. While the perceptual features from VGG already work well, the original VGG architecture can be further improved. Radford et. al [30] reported that as far as leaky ReLU avoids sparse gradients, replacing ReLUs with leaky ReLUs [17] in the discriminator stabilizes the training process of GANs. For the same reasons, changing max pooling layers to average pooling removes unwanted sparseness in the backpropagated gradients. Following these observations, we construct the *VGG** network, which is particularly suitable for the adversarial game. We thus took the VGG-19 network pretrained on ILSVRC dataset, replaced all max pooling layers by average poolings, ReLU nonlinearities by leaky ReLUs with a negative slope 0.2 and then trained on the ILSVRC dataset for further two days. We compare the variants of our approach based on VGG and VGG* features below.

Generator Architecture. For the image manipulation experiments, we used transformer network proposed by Johnson et al. [21]. It consists of M convolutional layers with stride size 2, N residual blocks [16] and M upsampling layers, each one increases resolution by a factor of 2. We set M and N in a way that allows outputs of the last residual block to have large enough receptive field, but at the same time for generator and discriminator to have similar number of parameters. We provide detailed descriptions of architectures in [2].

Stabilizing the Generator. We have also used two additional methods to improve the generator learning and to prevent its collapse. First, we have added the *identity loss* [36,42] that ensures that the generator does not change the input, when it comes from the p_{target}. Thus, the following term is added to the

maximization objective of the generator:

$$J_{\text{id}}^{G} = -\lambda_{\text{id}}\, \mathbb{E}_{\mathbf{y}\sim p_{\text{target}}} \lambda \left\| \mathbf{y} - G(\mathbf{y}) \right\|_{L_1}, \tag{7}$$

where λ_{id} is a meta-parameter that controls the contribution of the weight, and $\|\cdot\|_{L_1}$ denotes pixel-wise L1-metric.

To achieve the best results for the hardest translation tasks, we have found the cycle idea from the CycleGAN [42] needed. We thus train two generators $G_{\mathbf{x}\to\mathbf{y}}$ and $G_{\mathbf{y}\to\mathbf{x}}$ operating in opposite directions in parallel (and jointly with two discriminators), while adding reciprocity terms ensuring that mappings $G_{\mathbf{x}\to\mathbf{y}} \circ G_{\mathbf{y}\to\mathbf{x}}$ and $G_{\mathbf{y}\to\mathbf{x}} \circ G_{\mathbf{x}\to\mathbf{y}}$ are close to identity mappings.

Moreover, we notice that usage of external features as inputs for the discriminator leads to fast convergence of the discriminator loss to zero. Even though this is expected, since our method essentially corresponds to pretraining of the discriminator, this behavior is one of the GAN failure cases [8] and on practice leads to bad results in harder tasks. Therefore we find pretraining of the generator to be required for increased stability. For image translation task we pretrain generator as autoencoder. Moreover, the necessity to pretrain the generator makes our approach fail to operate in DCGAN setting with unconditional generator.

After an additional stabilization through the pretraining and the identity and/or cycle losses, the generator becomes less prone to collapse. Overall, in the resulting approach it is neither easy for the discriminator to overpower the generator (this is prevented by the identity and/or cycle losses), nor is it easy for the generator to overpower the discriminator (as the latter always has access to perceptual features, which are good at judging the realism of the output).

4 Experiments

The goal of the experimental validation is two-fold. The primary goal is to validate the effect of perceptual discriminators as compared to baseline architectures which use traditional discriminators that do not have access to perceptual features. The secondary goal is to validate the ability of our full system based on perceptual discriminators to handle harder image translation/manipulation task with higher resolution and with less data. Extensive additional results are available on our project page [2]. We perform the bulk of our experiments on CelebA dataset [25], due to its large size, popularity and the availability of the attribute annotations (the dataset comprises over 200k of roughly-aligned images with 40 binary attributes; we use 160×160 central crops of the images). As harder image translation task, we use CelebA-HQ [22] dataset, which consists of high resolution versions of images from CelebA and is smaller in size. Lastly, we evaluate our model on problems with non-face datasets like apples to oranges and photo to Monet texture transfer tasks.

Experiments were carried out on NVIDIA DGX-2 server.

Fig. 2: Qualitative comparison of the proposed systems as well as baselines for neutral→smile image manipulation. As baselines, we show the results of DFI (perceptual features, no adversarial training) and DCGAN (same generator, no perceptual features in the discriminator). Systems with perceptual discriminators output more plausible manipulations.

Qualitative Comparison on CelebA. Even though our contribution is orthogonal to a particular GAN-based image translation method, we chose one of them, provided modifications we proposed and compared it with the following important baselines in an attribute manipulation task:

- *DCGAN* [30]: in this baseline GAN system we used image translation model with generator and discriminator trained only with adversarial loss.
- *CycleGAN* [42]: this GAN-based method learns two reciprocal transforms in parallel with two discriminators in two domains. We have used the authors' code (PyTorch version).
- *DFI* [39]: to transform an image, this approach first determines target VGG feature representation by adding the feature vector corresponding to input image and the shift vector calculated using nearest neighbours in both domains. Then the resulting image is produced using optimization-based feature inversion as in [28]. We have used the authors' code.
- *FaceApp* [1]: is a very popular closed-source app that is known for the quality of its filters (transforms), although the exact algorithmic details are unknown.

Our model is represented by two basic variants.

- *VGG-GAN*: we use DCGAN as our base model. The discriminator has a single classifier and no generator pretraining or regularization is applied, other than identity loss mentioned in the previous section.
- *VGG*-GAN*: same as the previous model, but we use a finetuned VGG network variant with dense gradients.

Table 1: Quantitative comparison: (a) Photorealism user study. We show the fraction of times each method has been chosen as "the best" among all in terms of photorealism and identity preservation (the higher the better). (b) C2ST results (cross-entropy, the higher the better). (c) Log-loss of classifier trained on real data for each class (the lower the better). See main text for details.

	(a) User study		(b) C2ST, $\times 10^{-2}$			(c) Classification loss		
	Smile	Age	Smile	Gender	Hair color	Smile	Gender	Hair color
DFI [39]	0.16	0.4	< 0.1	< 0.01	< 0.01	1.3	0.5	1.14
FaceApp [1]	0.45	0.41	–	–	–	–	–	–
DCGAN [30]	–	–	0.6	0.03	0.06	0.6	1.5	2.33
CycleGAN [42]	0.03	0.04	5.3	0.35	0.49	1.2	0.8	2.41
VGG-GAN	–	–	8.6	0.21	0.96	0.4	0.1	1.3
VGG*-GAN	0.36	0.15	5.2	0.24	1.29	0.7	0.1	1.24
Real data	–	–	–	–	–	0.1	0.01	0.56

The comparison with state-of-the-art image transformation systems is performed to verify the competitiveness of the proposed architecture (Figure 2). In general, we observe that VGG*-GAN and VGG-GAN models consistently outperformed DCGAN variant, achieving higher effective resolution and obtaining more plausible high-frequency details in the resulting images. While a more complex CycleGAN system is also capable of generating crisp images, we found that the synthesized smile often does not look plausible and does not match the face. DFI turns out to be successful in attribute manipulation, yet often produces undesirable artifacts, while FaceApp shows photorealistic results, but with low attribute diversity. Here we also evaluate the contribution of dense gradients idea for VGG encoder and find it providing minor quality improvements.

User Photorealism Study on CelebA. We have also performed an informal user study of the photorealism. The study enrolled 30 subjects unrelated to computer vision and evaluated the photorealism of VGG*-GAN, DFI, CycleGAN and FaceApp on smile and aging/rejuvenation transforms. To assess the photorealism, the subjects were presented quintuplets of photographs unseen during training. In each quintuplet the center photo was an image without the target attribute (e. g. real photo of neutral expression), while the other four pictures were manipulated by one of the methods and presented in random order. The subjects were then asked to pick one of the four manipulations that they found most plausible (both in terms of realism and identity preservation). While there was no hard time limit, the users were asked to make the pick as quickly as possible. Each subject was presented overall 30 quintuplets with 15 quantuplets allocated for each of the considered attribute. The results in Table 1a show that VGG*-GAN is competitive and in particular considerably better than the other feed-forward method in the comparison (CycleGAN), but FaceApp being the

Fig. 3: Results for VGG*-MS-CycleGAN attribute editing at 256×256 resolution on Celeba-HQ dataset. Networks have been trained to perform pairwise domain translation between the values of hair color, gender and smile attributes. Digital zoom-in recommended. See [2] for more manipulation examples.

winner overall. This comes with the caveat that the training set of FaceApp is likely to be bigger than CelebA. We also speculate that the diversity of smiles in FaceApp seems to be lower (Figure 2), which is the deficiency that is not reflected in this user study.

Quantitative Results on CelebA. To get objective performance measure, we have used the classifier two-sample test (C2ST) [26] to quantitatively compare GANs with the proposed discriminators to other methods. For each method, we have thus learned a separate classifier to discriminate between hold-out set of real images from target distribution and synthesized images, produced by each of the methods. We split both hold-out set and the set of fake images into training and testing parts, fit the classifier to the training set and report the log-loss over the testing set in the Table 1b. The results comply with the qualitative observations: artifacts, produced by DCGAN and DFI are being easily detected by the classifier resulting in a very low log-loss. The proposed system stays on par with a more complex CycleGAN (better on two transforms out of three), proving that a perceptual discriminator can remove the need in two additional networks and cycle losses. Additionally, we evaluated attribute translation performance

(a) (b) (c) (d) (e) (a) (b) (c) (d) (e)

Fig. 4: We compare different architectures for the discriminator on CelebA-HQ 256×256 male $\leftrightarrow$ female problem. We train all architectures in CycleGAN manner with LSGAN objective and compare different discriminator architectures. (a) Input, (b) VGG*-MS-CycleGAN: multi-scale perceptual discriminator with pretrained VGG* as a feature network F, (c) Rand-MS-CycleGAN: multi-scale perceptual discriminator with a feature network F having VGG* architecture with randomly-initialized weights, (d) MS-CycleGAN: multi-scale discriminator with the trunk shared across scales (as in our framework), where images serve as a direct input, (e) separate multi-scale discriminators similar to Wang et al. [40]. Digital zoom-in recommended.

in a similar fashion to StarGAN [9]. We have trained a model for attribute classification on CelebA and measured average log-likelihood for the synthetic and real data to belong to the target class. Our method achieved lower log-loss than other methods on two out of three face attributes (see Table 1c).

Higher Resolution. We further evaluate our model on CelebA-HQ dataset. Here in order to obtain high quality results we use all proposed regularization methods. We refer to our best model as VGG*-MS-CycleGAN, which corresponds to the usage of VGG* network with dense gradients as an encoder, multi-scale perceptual discriminator based on VGG* network, CycleGAN regularization and pretraining of the generator. Following CycleGAN, we use LSGAN [29] as an adversarial objective for that model. We trained on 256×256 version of CelebA-HQ dataset and present attribute manipulation results in Figure 3. As we can see, our model provides photorealistic samples while capturing differences between the attributes even for smaller amount of training samples (few thousands per domain) and higher resolution compared to our previous tests.

In order to ensure that each of our individual contributions affects the quality of these results, we consider three variations of our discriminator architecture and compare them to the alternative multi-scale discriminator proposed in Wang et al. [40]. While Wang et al. used multiple identical discriminators operating at different scales, we argue that this architecture has redundancy in terms of number of parameters and can be reduced to our architecture by combining these discriminators into a single network with shared trunk and separate multi-scale output branches (as is done in our method). Both variants are included into the comparison in Figure 4. Also we consider *Rand-MS-CycleGAN* baseline that uses random weights in the feature extractor in order to tease apart the

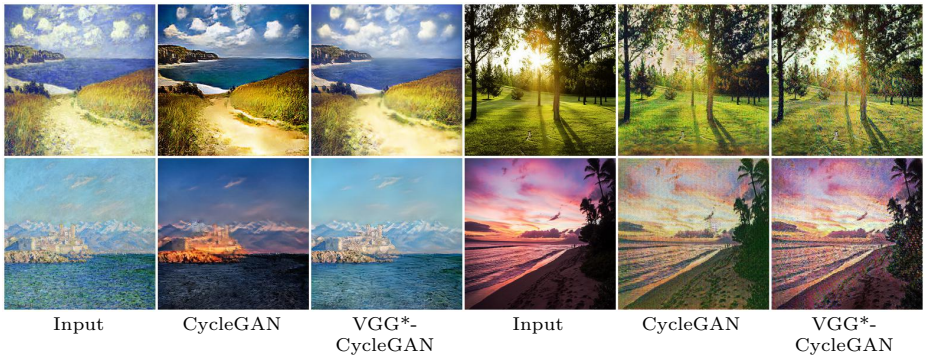

Fig. 5: Comparison between CycleGAN and VGG*-MS-CycleGAN on painting↔photo translation task. It demonstrates the applicability of our approach beyond face image manipulation. See [2] for more examples.

contribution of VGG* architecture as a feature network F and the effect of also having its weights pretrained on the success of the adversarial training. While the weights inside the VGG part were not frozen, so that adversarial training process could theoretically evolve good features in the discriminator, we were unable to make this baseline produce reasonable results. For high weight of the identity loss λ_{id} the resulting generator network produced near-identical results to the inputs, while decreasing λ_{id} lead to severe generator collapse. We conclude that the architecture alone cannot explain the good performance of perceptual discriminators (which is validated below) and that having pretrained weights in the feature network is important.

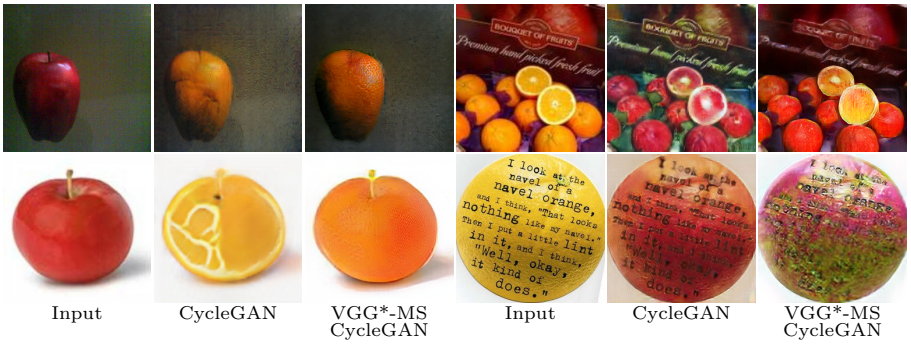

Fig. 6: Apple↔orange translation samples with CycleGAN and VGG*-MS-CycleGAN are shown. Zoom-in recommended. See [2] for more examples.

Non-face Datasets. While the focus of our evaluation was on face attribute modification tasks, our contribution applies to other translation tasks, as we verify in this section by performing qualitative comparison with the CycleGAN and VGG*-MS-CycleGAN architectures on two non-face domains on which CycleGAN was originally evaluated: an artistic style transfer task (Monet-photographs) in Figure 5 and an apple-orange conversion in Figure 6 (the figures show representative results). To achieve fair comparison, we use the same amount of residual blocks and channels in the generator and the same number of downsampling layers and initial amount of channels in discriminator both in our model and in the original CycleGAN. We used the authors' implementation of CycleGAN with default parameters. While the results on the style transfer task are inconclusive, for the harder apple-to-orange task we generally observe the performance of perceptual discriminators to be better.

Other Learning Formulations. Above, we have provided the evaluation of the perceptual discriminator idea to unaligned image translation tasks. In principle, perceptual discriminators can be used for other tasks, e.g. for unconditional generation and aligned image translation. In our preliminary experiments, we however were not able to achieve improvement over properly tuned baselines. In particular, for aligned image translation (including image superresolution) an additive combination of standard discriminator architectures and perceptual losses performs just as well as our method. This is not surprising, since the presence of alignment means that perceptual losses can be computed straight-forwardly, while they also stabilize the GAN learning in this case. For unconditional image generation, a naive application of our idea leads to discriminators that quickly overpower generators in the initial stages of the game leading to learning collapse.

5 Summary

We have presented a new discriminator architecture for adversarial training that incorporates perceptual loss ideas with adversarial training. We have demonstrated its usefulness for unaligned image translation tasks, where the direct application of perceptual losses is infeasible. Our approach can be regarded as an instance of a more general idea of using transfer learning, so that easier discriminative learning formulations can be used to stabilize and improve GANs and other generative learning formulations.

Acknowledgements. This work has been supported by the Ministry of Education and Science of the Russian Federation (grant 14.756.31.0001).

References

1. Faceapp. https://www.faceapp.com/ (2018)
2. Project webpage. http://egorzakharov.github.io/perceptual_gan (2018)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Proc. ICML. pp. 214–223 (2017)
4. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.S.: Neural codes for image retrieval. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. pp. 584–599 (2014)
5. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. In: Proc. NIPS. pp. 752–762 (2017)
6. Brock, A., Lim, T., Ritchie, J.M., Weston, N.: Neural photo editing with introspective adversarial networks. CoRR **abs/1609.07093** (2016)
7. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: Proc. ICCV. pp. 1520–1529 (2017)
8. Chintala, S., Denton, E., Arjovsky, M., Mathieu, M.: How to train a GAN? Tips and tricks to make GANs work. https://github.com/soumith/ganhacks (2017)
9. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proc. CVPR (2018)
10. Dong, C., Loy, C.C., He, K., Tang, X.: Learning a deep convolutional network for image super-resolution. In: Proc. ECCV. pp. 184–199 (2014)
11. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Proc. NIPS. pp. 658–666 (2016)
12. Dosovitskiy, A., Springenberg, J.T., Brox, T.: Learning to generate chairs with convolutional neural networks. In: Proc. CVPR. pp. 1538–1546 (2015)
13. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. CVPR. pp. 2414–2423 (2016)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proc. NIPS. pp. 2672–2680 (2014)
15. Goodfellow, I.J.: NIPS 2016 tutorial: Generative adversarial networks. CoRR **abs/1701.00160** (2017)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015), http://arxiv.org/abs/1512.03385
17. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 1026–1034 (2015)
18. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. **36**(4), 107:1–107:14 (2017)
19. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proc. CVPR. pp. 5967–5976 (2017)
20. Jain, V., Seung, S.: Natural image denoising with convolutional networks. In: Proc. NIPS. pp. 769–776 (2009)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Proc. ECCV. pp. 694–711 (2016)
22. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. CoRR **abs/1710.10196** (2017)

23. Kim, J., Kwon Lee, J., Mu Lee, K.: Accurate image super-resolution using very deep convolutional networks. In: Proc. CVPR. pp. 1646–1654 (2016)
24. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Proc. CVPR (2017)
25. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proc. ICCV (2015)
26. Lopez-Paz, D., Oquab, M.: Revisiting classifier two-sample tests. arXiv preprint arXiv:1610.06545 (2016)
27. Lucic, M., Kurach, K., Michalski, M., Gelly, S., Bousquet, O.: Are GANs created equal? A large-scale study. CoRR **abs/1711.10337** (2017)
28. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proc. CVPR (2015)
29. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z.: Multi-class generative adversarial networks with the L2 loss function. CoRR **abs/1611.04076** (2016)
30. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR **abs/1511.06434** (2015)
31. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: CNN features off-the-shelf: An astounding baseline for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 2014. pp. 512–519 (2014)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. CoRR **abs/1409.0575** (2014), `http://arxiv.org/abs/1409.0575`
33. Sajjadi, M.S.M., Scholkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: Proc. ICCV (2017)
34. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Proc. NIPS. pp. 2226–2234 (2016)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
36. Taigman, Y., Polyak, A., Wolf, L.: Unsupervised cross-domain image generation. CoRR **abs/1611.02200** (2016), `http://arxiv.org/abs/1611.02200`
37. Ulyanov, D., Lebedev, V., Vedaldi, A., Lempitsky, V.S.: Texture networks: Feed-forward synthesis of textures and stylized images. In: Proc. ICML. pp. 1349–1357 (2016)
38. Ulyanov, D., Vedaldi, A., Lempitsky, V.S.: Deep image prior. In: Proc. CVPR (2018)
39. Upchurch, P., Gardner, J.R., Pleiss, G., Pless, R., Snavely, N., Bala, K., Weinberger, K.Q.: Deep feature interpolation for image content changes. In: Proc. CVPR. pp. 6090–6099 (2017)
40. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. arXiv preprint arXiv:1711.11585 (2017)
41. Zhang, H., Xu, T., Li, H., Zhang, S., Huang, X., Wang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. CoRR **abs/1612.03242** (2016)
42. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. ICCV. pp. 2242–2251 (2017)

43. Zhu, J., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Proc. NIPS. pp. 465–476 (2017)