Discriminative Region Proposal Adversarial Networks for High-Quality Image-to-Image Translation

Chao Wang[0000−0002−1177−8104], Haiyong Zheng*[0000−0002−8027−0734], Zhibin Yu[0000−0003−4372−1767], Ziqiang Zheng[0000−0003−4027−7955], Zhaorui Gu[0000−0002−6673−7932], and Bing Zheng[0000−0003−2295−3569]

Ocean University of China, Qingdao 266100, China
chaowangplus@gmail.com, {zhenghaiyong, yuzhibin}@ouc.edu.cn,
zhengziqiang@stu.ouc.edu.cn, {guzhaorui, bingzh}@ouc.edu.cn
http://vision.ouc.edu.cn
* Corresponding author

Abstract. Image-to-image translation has been made much progress with embracing Generative Adversarial Networks (GANs). However, it's still very challenging for translation tasks that require high quality, especially at high-resolution and photorealism. In this paper, we present Discriminative Region Proposal Adversarial Networks (DRPAN) for high-quality image-to-image translation. We decompose the procedure of image-to-image translation task into three iterated steps, first is to generate an image with global structure but some local artifacts (via GAN), second is using our DRPnet to propose the most fake region from the generated image, and third is to implement "image inpainting" on the most fake region for more realistic result through a reviser, so that the system (DRPAN) can be gradually optimized to synthesize images with more attention on the most artifact local part. Experiments on a variety of image-to-image translation tasks and datasets validate that our method outperforms state-of-the-arts for producing high-quality translation results in terms of both human perceptual studies and automatic quantitative measures.

Keywords: GAN · DRPAN · Image-to-image translation.

1 Introduction

From the aspect of human visual perception, why we consider a synthesized image as fake is often because it contains local artifacts. Although it looks like real at the first glance, we can still easily distinguish the fake from the real by gazing for only about $1000ms$ [5]. Human being has the ability to draw a realistic scene from coarse structure to fine detail, that is, we usually get the global structure of a scene while focus on the detail of an object and understand how it is associated with surroundings. Under this intuition, our goal of this work is to develop an image-to-image translation system for high-quality image synthesis with clear structure and vivid details.

Many efforts have been made to develop an automatic image-to-image translation system. The straightforward approach was to optimize on pixel-wise space with L1 or L2 loss [9,23]. However, both of them suffer from blur problem. So some works added adversarial loss for generating more sharp images in both spatial and spectral dimensions [14]. Except for the GAN loss, perceptual loss has been used in image-to-image translation tasks, but it was limited to a pretraining deep model and the training datasets [37]. Although we have a variety of losses to evaluate the discrepancy between real image and generated image, using GAN for image-to-image translation still encounters with the artifacts and unsmooth color distribution problems, and it is even hard to generate high-resolution photo-realistic images because of the high dimension distribution [26].

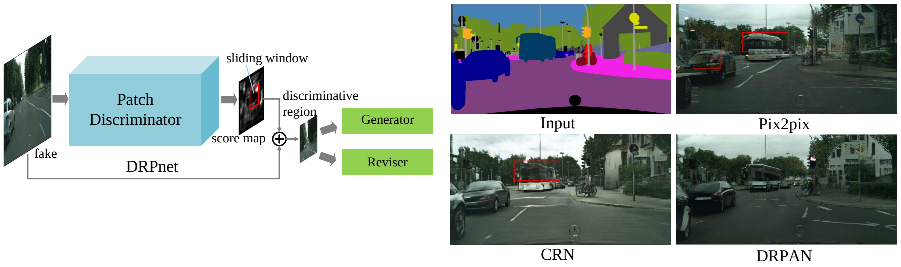

Fig. 1. Left: Our Discriminative Region Proposal network (DRPnet). **Right**: Synthesized samples compared with previous works on Cityscapes validation dataset [6]. The regions within red window show obvious artifacts or deformation. Our method can synthesize images with clear structure and vivid details.

So, how could we solve this problem intuitively? We decompose the procedure of image-to-image translation task into three iterated steps, first is to generate an image with global structure but some local artifacts (via GAN), second is to propose the most fake region from the generated image (using our DRPnet shown in Fig. 1), and third is to implement "image inpainting" on the most fake region for more realistic result, so that the system (our DRPAN) can be gradually optimized to synthesize images with more attention on the most artifact local part. Inspired by this motivation, we develop a framework based on patch-wise discriminator to predict the discriminative score map and use sliding windows to find the most artificial region. Then the proposed discriminative region will be used to mask the corresponding real sample and output as "masked fake". Finally, we propose a reviser to distinguish the real from the masked fake for producing realistic details and serve as auxiliaries for generator to synthesize high-quality translation results. The reviser will critic on the fake image iteratively with different regions. We provide a weighted parameter to balance the contribution of the patch discriminator and our reviser for different levels of translation tasks. Using this proposed DRPAN, we can synthesize high-quality images with high-resolution and photo-reality details but less artifacts.

The main contribution of the study is threefold: first, we design the mechanism to explore patch-based discriminators for producing discriminative region; second, we propose the reviser for GANs to provide constructive revisions for generator which usually are missed by patch discriminator; third, we build a DRPAN model as a general-purpose solution for high-quality image-to-image translation tasks on different levels. The code of this paper is available at https://github.com/godisboy/DRPAN.

2 Related works

Feed-forward based approach. Deep Convolutional Neural Networks (CNNs) have been performed well on many computer vision tasks. For style transform problems [15], many studies were mainly based on VGG-16 network architecture [20] and used perceptual losses for style translation [10]. Network architectures that work well on object recognition tasks have been proved to work well on generative models, *e.g.*, some computer vision translation and editing tasks used residual block as a strong feature learning representation architecture [19,22]. Feed-forward CNNs accompanied with per-pixel loss have been presented for image super-resolution [9,16,34,15], image colorization [8,44], and semantic segmentation [23,4,31]. A recent work for photo realistic image synthesis system, called CRN [5], can synthesize images with high resolution. However, the images synthesized by feed-forward based approach usually become smooth too much rather than realistic, *i.e.*, not sharp enough in details. Besides, these methods are limited to be applied to other image-to-image translation tasks.

GAN based approach. GANs [11] introduced an unsupervised method to learn real data distribution. And DCGAN [29] firstly used CNNs to train generative adversarial networks which was hard to be deployed in other tasks before. Then, CNNs were extensively used for designing GAN architectures. Towards stable training of GAN, WGAN [1] replaced Jensen-Shannon divergence by Wasserstein distance as the optimization metric, and recently a variety of more stable alternatives have been proposed [28,18,12]. Wang and Gupta [38] combined structured GAN with style GAN to learn to generate natural indoor scenes. Reed *et al.* [30] used text as conditional input to synthesize images with semantic variation. Pathak *et al.* [27] proposed context encoders for image inpainting accompanied by adversarial loss. Li *et al.* [21] trained GANs with a combination of reconstruction loss, two adversarial losses and a semantic parsing loss for face completion. Nguyen *et al.* [25] presented Plus and Play Generative Networks for high-resolution and photo-realistic image generation with the resolution of 227 × 227 images. Isola *et al.* explored [14] conditional GANs for a variety of image-to-image translation problems. ID-CGAN [43] combined conditional GANs with perceptual loss for single image de-raining and de-snowing. Considering that the paired images are less and hard to collect, some works proposed unpaired or unsupervised translation frameworks [46,17,40]. But it limits to the similarity of translation between source domain A and target domain B.

PatchGAN was firstly used in neural style transfer with CNNs based on patch feature inputs [20]. Pix2pix [14] showed that a full ImageGAN does not show quality improvement compared with a low 70×70 patch discriminator which has less parameters and needs low computing resource. SimGAN [35] used patch based score map for real image synthesis tasks and mapped a full image to a probability map. Our method explores PatchGAN to a unified discriminative region proposal network model for deciding where and how to synthesize via a reviser. We show that this approach can improve translation results on high-quality, especially at high-resolution and photo-reality.

3 Method

Our image-to-image translation model, called Discriminative Region Proposal Adversarial Networks (DRPAN), is composed of three components: a generator, a discriminator, and a reviser. The discriminator explores PatchGAN to construct Discriminative Region Proposal network (DRPnet, see Fig. 1) to find and extract the discriminative region for producing masked fake sample, while the reviser adopts CNN to distinguish the real from the masked fake to provide constructive revisions for generator. The overall network architecture and data flow are illustrated in Fig. 2.

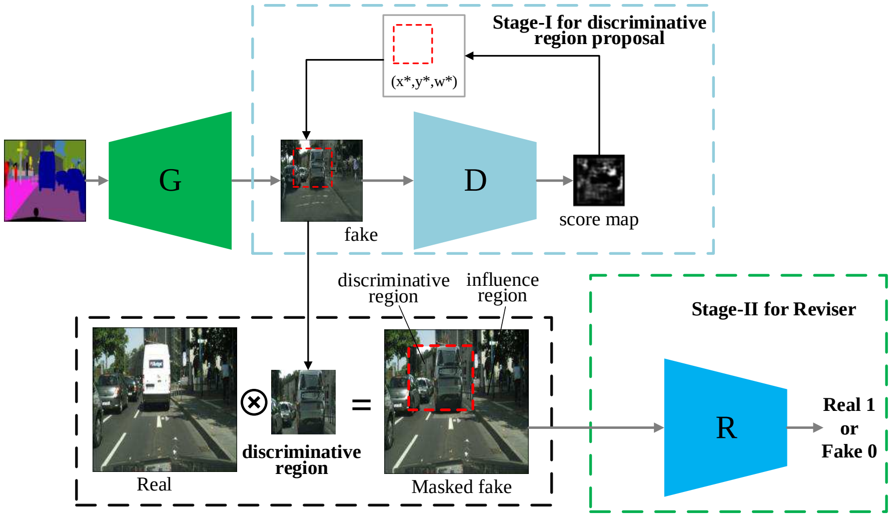

Fig. 2. The overall network architecture and data flow of our proposed Discriminative Region Proposal Adversarial Networks (DRPAN), which is composed of three components: a generator, a discriminator, and a reviser, and is a unified model for image-to-image translation tasks.

Fig. 3 shows our process of how to improve the quality of synthesized image. It can be seen that, as our DRPAN continues to train, the discriminative region for masked fake images (right) varies so that the quality of synthesized images (left) are improved with brighter score map (the first and the last). Besides, although it is hard to distinguish the synthesized sample from the real sample after many epochs, our DRPAN can still revise the generator to optimize the synthesized result in the details for high quality.

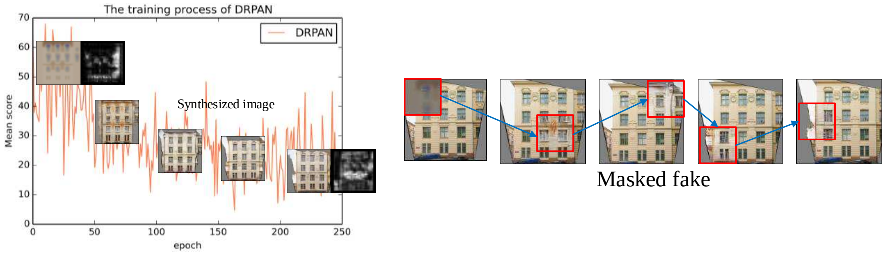

Fig. 3. The training process of DRPAN on facades dataset [36]. **Left**: The plotting curve shows mean value of score map on synthesized samples. **Right**: Step by step synthesis on different discriminative regions.

3.1 DRPAN

We first suggest that patch-based discriminators produce meaningful score maps, which may have applications beyond image synthesis. Fig. 4 shows the output results of score map on different quality levels (fake and real) of images by a pre-trained PatchGAN. It can be seen that, the score maps of the fake samples, which have obvious artifacts and shape deformation on some regions, are almost dark with lower score on the corresponding regions; in contrast, the score maps of the real samples are brightest with the highest scores. From the visualization of score maps, we can find the darkest region for proposing the discriminative region that indicates the remarkable fake region.

Based on the observation shown in Fig. 4, we explore patch discriminator to DRPnet for producing discriminative region. Given an input image with resolution $w_i \times w_i$, and it is processed by the patch discriminator to be a probability score map with size $w_s \times w_s$. Suppose we want to obtain the discriminative region at $w^* \times w^*$, the size of sliding window w for score map can be calculated by

$$w = w^* \times w_s / w_i. \tag{1}$$

Then our DRPnet will find the discriminative square patch on score map with the center coordinates (x_c, y_c) and length w, so the scale τ between the input image and output score map is

$$\tau = \frac{w_i - w*}{w_s - w}. \tag{2}$$

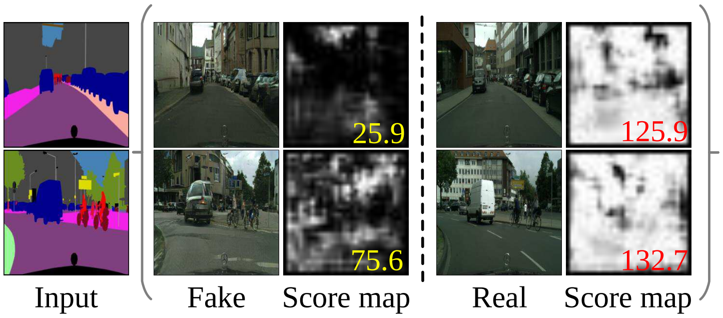

Fig. 4. The output results of score map on different quality levels (fake and real) of images by a pre-trained PatchGAN. The darkest regions on score maps mean the lowest quality, indicating that patch-based discriminators can be explored for discriminative region proposal.

The center coordinates (x_c^*, y_c^*) of discriminative region will be calculated by

$$\begin{cases} x_c^* = \tau \times x_c, \\ y_c^* = \tau \times y_c. \end{cases} \tag{3}$$

Finally, the discriminative region d_r produced by DRPnet can be expressed as

$$d_r = F_{\text{DRPnet}}(x_c^*, y_c^*, w^*). \tag{4}$$

Instead of only optimizing the independent local regions, we consider the relationship between fake discriminative region and real surrounding influence regions, so that it can connect the fake to the real for providing constructive revisions to generator. The influence region is defined as the region which is connected to the most fake regions and has semantic and spatial relationship with the content in it (e.g., the wheel is often below the car window). For this purpose, we mask the corresponding real sample using the fake discriminative region to make masked fake sample, and then design a reviser using CNN to distinguish real from masked fake to optimize the generator for synthesizing high-quality images. The reviser we proposed can also be used for other GANs to improve the quality of generated samples.

3.2 Objective

For image-to-image translation tasks, we not only want to generate the realistic samples, but also desire diversity with different conditional inputs. The original GANs suffer from unstability and mode collapse problems [1,2]. So some recent works [1,28,12] improved the training of GAN. To stably train our DRPAN with high-diversity synthesis ability, we modify DRAGAN [18] as the loss of our reviser R, and use the original objective function for training Patch Discriminator.

$$\mathcal{L}_D(G, D_p) = \mathbb{E}_y[\log D_p(x, y)] + \mathbb{E}_{x,z}[\log(1 - D_p(x, G(x, z)))]. \tag{5}$$

For reviser R, to distinguish between the very similar real and masked fake $y_{\text{mask}} = M(G(x,z))$ ($M(\cdot)$ represents the mask operation), we add a regularization to the loss of reviser as the penalty, which is expressed as

$$\mathcal{L}_R(G,R) = \mathbb{E}_y[\log R(x,y)] + \mathbb{E}_{x,z}[\log(1-R(x,y_{\text{mask}}))] + \alpha\mathbb{E}_{x,\delta}[\|\nabla_x R(x+\delta)\|-1]^2, \tag{6}$$

where α is hyper parameter, δ is random noise on x, and ∇ indicates gradient.

Previous studies have found it beneficial to mix the GAN objective with a more traditional loss, such as L2 and L1 distance [14,35]. Considering that L1 distance encourages less blurring than L2 [14], we provide extra L1 loss for regularization on the whole input image and the local discriminative region to generator, which is defined as

$$\mathcal{L}_{L_1}(G) = \beta\mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1] + \gamma\mathbb{E}_{d_r,y_r,z}[\|y_r - F_{\text{DRPnet}}(G(x,z))\|_1], \tag{7}$$

where β and γ are hyper parameters, d_r is the discriminative region, and y_r represents the region on the real image corresponding to the discriminative region on the synthesized image. Then the total loss of generator can be expressed as

$$\mathcal{L}_G(G,D_p,R) = -\mathbb{E}_{x,z}[\log(1-D_p(x,G(x,z)))] - \mathbb{E}_{x,z}[\log(1-R(x,y_{\text{mask}}))] + \mathcal{L}_{L_1}(G). \tag{8}$$

Our proposed model totally contains a generator G, a patch discriminator D_p for DRPnet, and a reviser R. G will be optimized by D_p, R and L_1. And our full objective function is

$$L(G,D_p,R) = (1-\lambda)\mathcal{L}_D(G,D_p) + \lambda\mathcal{L}_R(G,R) + \mathcal{L}_{L_1}(G), \tag{9}$$

where λ is a hyper parameter to balance $\mathcal{L}_D$ and $\mathcal{L}_R$.

3.3 Network architecture

For our generator, we use architecture based on [19] which has convincing power for single image super-resolution. We adopt convolution and fractionally convolution blocks for down and up sampling respectively, and 9 residual blocks [46] for task learning. Each layer uses Batch Normalization [13] and ReLU [24] as activation function. For patch discriminator, we mainly implement with 70 × 70 PatchGAN [20,14]. The DRPAN reviser is a discriminator modified on DCGAN [29] that has a global view on the whole input. At the end of both discriminator and reviser, we adopt Sigmoid as activation function to output probability.

4 Experiments

To evaluate the performance of our proposed method on image-to-image translation tasks, we deploy a variety of experiments about different levels of translation tasks to compare our method with state-of-the-arts. And for different tasks, we also use different evaluation metrics including human perceptual studies and automatic quantitative measures.

4.1 Evaluation metrics

Image quality evaluation. PSNR, SSIM [39] and VIF [33] are some of the most popular evaluation metrics in low-level computer vision tasks such as deblurring, dehazing and image restoration. So for de-raining and aerial to maps tasks, we adopt PSNR, SSIM, VIF and RECO [3] to qualify the performance of results.

Image segmentation evaluation metrics. We use standard metrics from Cityscapes benchmark [6] to evaluate real to semantic labels task on Cityscapes dataset, including per-pixel accuracy, per-class accuracy, and Class IOU.

Amazon Mechanical Turk (AMT). AMT [14,46,40] is adopted in many tasks as a gold metric to evaluate how real the synthesized images, and we use it as evaluation metric for semantic labels to photo and maps to aerial tasks.

FCN-8s score. The intuition of using an off-the-shelf classifiers for automatic quantitative measurement is that if the generated images are realistic, classifiers trained on real images will be able to classify the synthesized image correctly as well [14]. We use the FCN-8s score [23] to evaluate semantic labels to real task on Cityscapes dataset. The FCN-8s model trained on Cityscapes segmentation tasks is taken from [14].

4.2 Why DRPAN?

To study the influence of DRPAN for revising synthesis and different situations of loss between proposed region and real region. We set an experiments which start from a pre-trained PatchGAN and continue for several training pipelines: continue training with PatchGAN; continue training with PatchGAN and L1 loss of discirminative and real region; continue training with PatchGAN and reviser.

We argue that the PatchD is efficient to discover the most fake or real region (Fig 4) from the image but is limited to improve these regions with fine details for that PatchD is hard to capture the high dimension distribution. In this case, we propose a DRPnet (explore the strength of PatchD) for discriminative region proposal and design a reviser to gradually remove visual artifacts, and thus reduce it to lower dimension estimation problem. This can be seen as a “top-down” procedure which is different from other gradually “bottom-up” image generation method [42]. Fig 5 shows the necessity of our proposed DRPAN for high-quality image-to-image translation, which illustrates that continue training PatchD is no help to reduce artifacts even with a L1 loss for balance, and DRPAN with only L1 loss can smooth the artifacts but not very sharp in details, while DRPAN with reviser exceeds the PatchD’s performance with less visual artifacts. The combination of reviser and L1 loss can reduce these artifacts ignored by PatchD. We also find that fake-mask operation can improve the fluency of whole image in certain samples (*e.g.*, the connection between door and wall). So DRPAN with fake-mask is implemented in the following experiments.

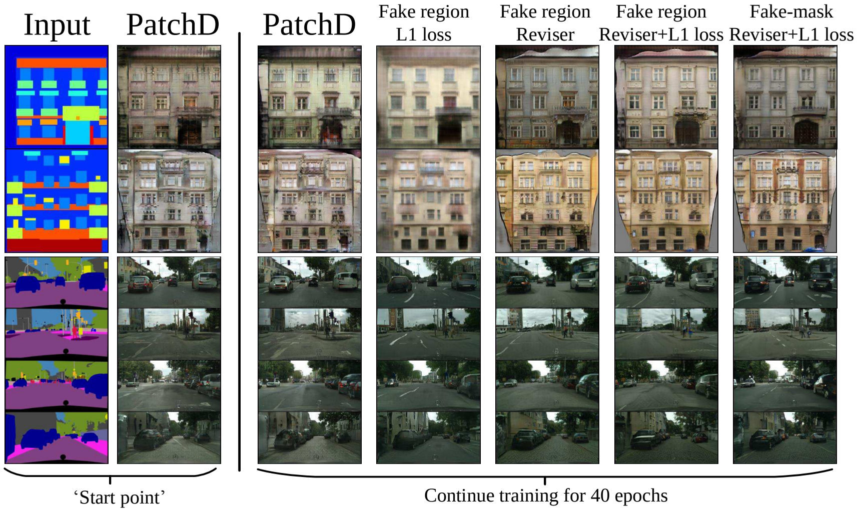
Input
PatchD
PatchD
Fake region L1 loss
Fake region Reviser
Fake region Reviser+L1 loss
Fake-mask Reviser+L1 loss
'Start point'
Continue training for 40 epochs

Fig. 5. Different methods with various losses produce different quality of results. The second column is the start point of comparison trained by PatchD, and all other models are continued trained for 40 epochs more. These experiments validate the necessary of our DRPnet for discriminative region proposal, our reviser for optimizing generator, and our fake-mask operation for improving synthesis.

4.3 Low level translation

We first apply our model on two low level translation tasks which are only related to the appearance translation of images, for example, in de-raining task we don't need change the content and texture of the input sample. So we set $\lambda = 1$ in Eqn. 9 for image synthesis using only reviser.

Single image de-raining. We trained and tested our DRPAN model on single image de-raining task using the procedure as same as [43], and evaluated the results by both qualitative and quantitative metrics. Fig. 6 shows the qualitative results of our DRPAN with different sizes of discriminative region compared to ID-CGAN [43], and DRPAN outperforms ID-CGAN with not only more effective de-raining but also more vivid color and clear details. Tab. 1 reports the corresponding quantitative results evaluated by PSNR, SSIM, VIF, and RECO metrics, and the best results (in bold font) are achieved all by our DRPAN.

Bw to color. We trained our DRPAN model for image colorization task on ImageNet [7], and tested on ImageNet val dataset with an example shown in Fig. 7. Our DRPAN can produce compelling colorization results compared with classification with class rebalancing [44]. In addition, we run AMT evaluation for colorization(Tab. 2). Our method fooled participants on 27.8% which is competitive with the full method from [44].

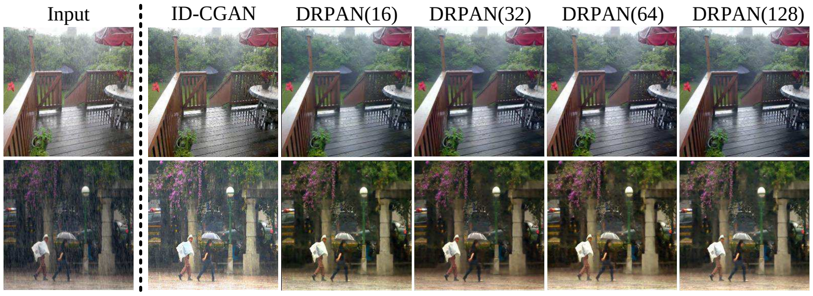

Fig. 6. Example results of our DRPAN with different sizes of discriminative region compared to ID-CGAN [43] on single image de-raining task.

Table 1. Quantitative comparison of our DRPAN (with different sizes of discriminative region) with ID-CGAN [43] and PAN [37] on image de-raining. DRPAN performs best (in bold font) evaluated by PSNR, SSIM, VIF, and RECO metrics

Method	L2+	ID-	PAN[37]	DRPAN	DRPAN	DRPAN	DRPAN	DRPAN
Metrics	CGAN	CGAN[43]		(w/o mask)	(128)	(64)	(32)	(16)
PSNR	22.19	22.91	23.35	25.51	25.87	25.76	25.92	**26.20**
SSIM	0.8083	0.8198	0.8303	0.8688	0.8714	0.8765	**0.8788**	0.8712
VIF	0.3640	0.3885	0.4050	0.4923	0.4818	0.4962	**0.5001**	0.4783
RECO	–	–	–	0.9670	1.0770	**1.1072**	1.1067	1.0875

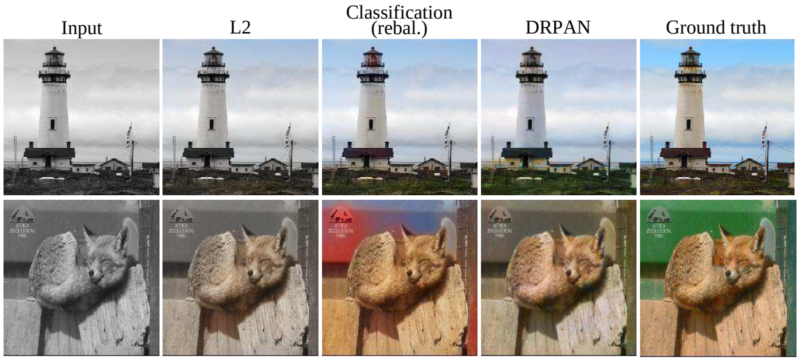

Fig. 7. Example results of our DRPAN compared to L2 regression [44] and Classification (rebal.) [44] on image colorization task.

4.4 Real to abstract translation

We then implement our proposed DRPAN on two tasks of real to abstract translation which requires many-to-one abstraction ability.

Real to semantic labels. For real to semantic labels task, we tested our DRPAN model on two of the most used datasets: Cityscapes and facades. Fig. 8

Table 2. AMT “real vs fake” test on corlorization

Method	% Turkers labeled real
L2 regression	23.4%
Classification	**29.7%**
DRPAN	27.8%

shows the qualitative results of our DRPAN compared to Pix2pix [14] on Cityscapes dataset for translating real to semantic labels, and DRPAN can synthesize more realistic results that are closer to ground truth than Pix2pix, meanwhile, the quantitative results in Tab. 3 can also tell this in terms of per-pixel accuracy, per-class accuracy, and Class IOU.

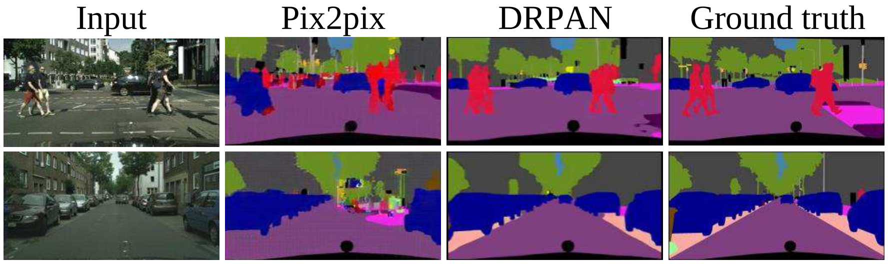

Fig. 8. Example results of our DRPAN compared to Pix2pix [14] on real to semantic labels task.

Aerial to maps. We also applied our DRPAN on aerial photo to maps task, and the experiment was implemented using paired images with 512 × 512 resolution [14]. The top row of Fig. 9 shows the qualitative results of our DRPAN compared to Pix2pix [14], indicating that our DRPAN can correctly translate the motorway on aerial photo into the orange line on the map while Pix2pix can't.

4.5 Abstract to real translation

Besides, we also demonstrate our proposed DRPAN on several abstract to real tasks that can translate one to many: semantic labels to photo, maps to aerial, edge to real, and sketch to real.

Semantic labels to real. For semantic labels to real task, the translation model aims to synthesize real world images from semantic labels. CGAN based works fail to capture the details in the real world and suffer from deformation and blur problems. CNN based methods such as CRN can synthesize high-resolution but smooth rather than realistic results. Fig. 10 shows qualitative comparison of results, from which it can be seen that our DRPAN can synthesize the most

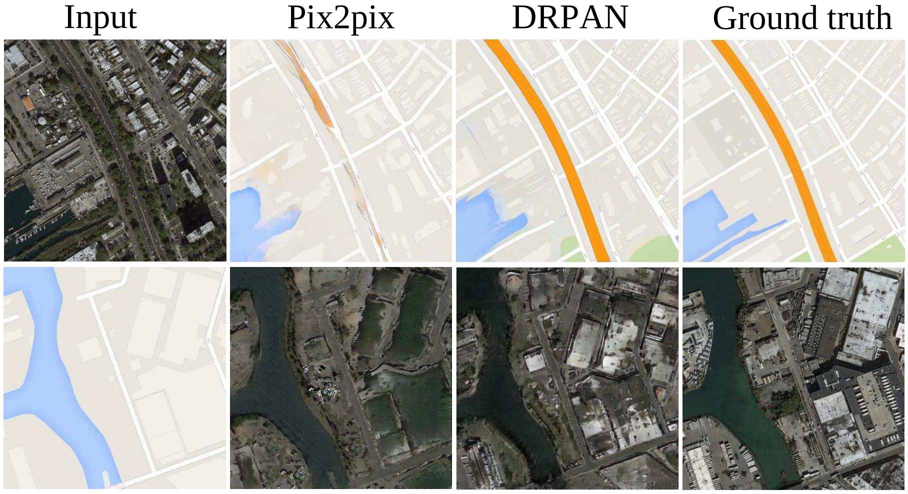

Fig. 9. Example results of our DRPAN compared to Pix2pix [14] on aerial to maps (top) and maps to aerial (bottom) tasks.

realistic results with high-quality (more clear and less distorted while high resolution) compared to Pix2pix [14] and CRN [5].

The evaluation of GAN is still a challenging problem. Many works [32,38,44,14] used off-the-shelf classifiers as automatic measures of synthesized images. Tab. 4 reports performance evaluation on segmentation of FCN-8s model, and our DRPAN exceeds Pix2pix [14] by 10% on per-pixel accuracy and also achieves highest performance on per-class accuracy and Class IOU.

Table 3. Quantitative comparison of our DRPAN with Pix2pix [14] on real to semantic labels task (Cityscapes dataset)

Model	Per-pixel acc.	Per-class acc.	Class IOU
L1+U-Net [14]	0.86	0.42	0.35
Pix2pix [14]	0.83	0.36	0.29
DRPAN(w/o fake-mask)	**0.86**	**0.48**	**0.39**
DRPAN	**0.88**	**0.52**	**0.43**

Table 4. Quantitative comparison of our DRPAN with other models on semantic labels to real task (Cityscapes dataset) by FCN-8s score

Model	Per-pixel acc.	Per-class acc.	Class IOU
L1+CGAN [14]	0.63	0.21	0.16
CRN	0.69	0.21	**0.20**
DRPAN(w/o fake-mask)	**0.72**	**0.22**	0.19
DRPAN	**0.73**	**0.24**	0.19
Ground truth	0.80	0.26	0.21

Maps to aerial. As opposed to aerial to maps task, we also tested our DRPAN on maps to aerial task, and the qualitative results are shown in the bottom row of Fig. 9, which clearly demonstrates that our DRPAN can synthesize higher quality aerial photos than Pix2pix [14].

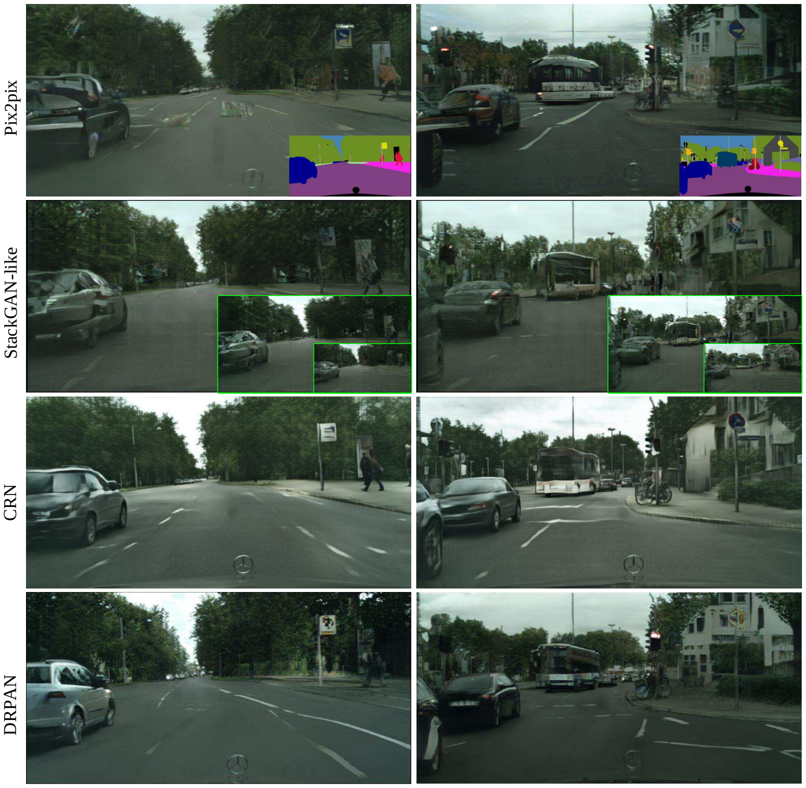

Fig. 10. Example results of our DRPAN compared to Pix2pix [14] and CRN [5] on semantic labels to real task with 512 × 512 resolution.

Table 5. AMT real vs. fake results test on Cityscapes semantic labels to photo task

Model	% Turkers labeled real
Pix2pix [14]	5.3%
StackGAN-like [42]	6.8%
CRN [5]	9.4%
DRPAN(w/o fake-mask)	**14.3%**
DRPAN	**18.2%**
	% Turkers labeled more realistic
DRPAN vs. Pix2pix [14]	**91.2%**
DRPAN vs. StackGAN-like	**84.6%**
DRPAN vs. CRN [5]	**75.7%**

Table 6. AMT real vs. fake results test on maps to aerial task

Model	% Turkers labeled real
Pix2pix [14]	25.2%
DRPAN(w/o fake-mask)	31.7%
DRPAN	**33.4%**

Human perceptual validation. We assess the performance of abstract to real on semantic labels to photo and maps to aerial by AMT. For fake against real study, we followed the perceptual study protocol from [14], and collected data of each algorithm from 30 participants. Each participant has 1000*ms* to look one sample. We also compared how realistic the synthesized images between different algorithms. Tab. 5 illustrates that images synthesized by DRPAN are ranked more realistic than state-of-the-arts (DRPAN 18.2% > CRN 9.4% > StackGAN-like 6.8% > Pix2pix 5.3%), moreover, compared to Pix2pix [14], StackGAN-like [42] and CRN [5], images synthesized by DRPAN are ranked more realistic by 91.2%, 84.6% and 75.7% respectively. Tab. 6 reports the comparison on maps to aerial task and our DRPAN fooled participants on 39.0% over 18.7% of Pix2pix and 26.8% of CycleGAN [46] respectively.

Edges to real and sketch to real. For the edge to real and sketch to real tasks, previous works often encounter with two problems [14]: one is that it's easy to generate artifacts and artificial color distribution in regions when the input such as edge is sparse; the other is that it's difficult to deal with unusual inputs like sketch. We tested our DRPAN model on UT Zappos50k dataset [41] and edge to handbag dataset [45]. Fig. 11 shows that our model can also handle these two problems well.

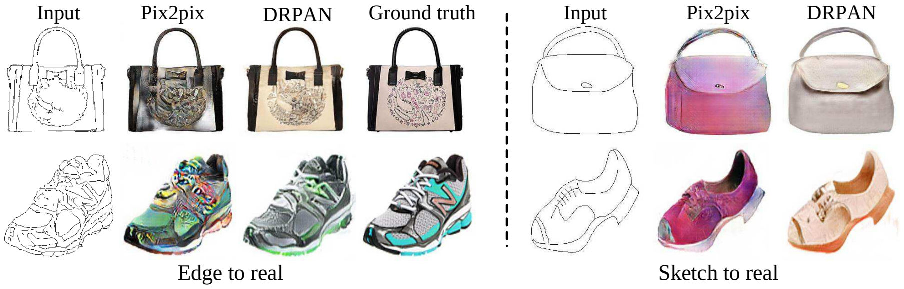

Fig. 11. Example results of our DRPAN compared to Pix2pix [14] on edge to real (left) and sketch to real (right) tasks.

5 Conclusions

We propose Discriminative Region Proposal Adversarial Networks (DRPAN) towards high-resolution and photo-reality image-to-image translation. Human perceptual studies and automatic quantitative measures validate the performance of our proposed DRPAN against the state-of-the-arts for synthesizing high-quality results. We hope it can be explored for discriminative feature learning and other computer vision tasks in the future.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 61771440 and 41776113, and Qingdao Municipal Science and Technology Program under Grant 17-1-1-5-jch.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein GAN. In: ICML (2017)
2. Arora, S., Ge, R., Liang, Y., Ma, T., Zhang, Y.: Generalization and equilibrium in generative adversarial nets (GANs). arXiv preprint arXiv:1703.00573 (2017)
3. Baroncini, V., Capodiferro, L., Di Claudio, E.D., Jacovitti, G.: The polar edge coherence: a quasi blind metric for video quality assessment. In: ESPC (2009)
4. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR (2015)
5. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV (2017)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR (2009)
8. Deshpande, A., Rock, J., Forsyth, D.: Learning large-scale automatic image colorization. In: ICCV (2015)
9. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE TPAMI **38**(2), 295–307 (2016)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
11. Goodfellow, I., Pougetabadie, J., Mirza, M., Xu, B., Wardefarley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NIPS (2014)
12. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.: Improved training of Wasserstein GANs. arXiv preprint arXiv:1704.00028 (2017)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
15. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: ECCV (2016)
16. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: CVPR (2016)
17. Kim, T., Cha, M., Kim, H., Lee, J., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. arXiv preprint arXiv:1703.05192 (2017)
18. Kodali, N., Abernethy, J., Hays, J., Kira, Z.: How to train your DRAGAN. arXiv preprint arXiv:1705.07215 (2017)
19. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
20. Li, C., Wand, M.: Precomputed real-time texture synthesis with Markovian generative adversarial networks. In: ECCV (2016)
21. Li, Y., Liu, S., Yang, J., Yang, M.H.: Generative face completion. In: CVPR (2017)
22. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (2017)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)

24. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: ICML (2010)
25. Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. In: CVPR (2017)
26. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier GANs. In: ICLR (2016)
27. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR (2016)
28. Qi, G.J.: Loss-sensitive generative adversarial networks on Lipschitz densities. arXiv preprint arXiv:1701.06264 (2017)
29. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
30. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. In: ICML (2016)
31. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
32. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NIPS (2016)
33. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE TIP **15**(2), 430–444 (2006)
34. Shi, W., Caballero, J., Huszr, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR (2016)
35. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: CVPR (2017)
36. Tyleček, R., Šára, R.: Spatial pattern templates for recognition of objects with regular structure. In: GCPR (2013)
37. Wang, C., Xu, C., Wang, C., Tao, D.: Perceptual adversarial networks for image-to-image transformation. In: IJCAI (2017)
38. Wang, X., Gupta, A.: Generative image modeling using style and structure adversarial networks. In: ECCV (2016)
39. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
40. Yi, Z., Zhang, H., Tan, P., Gong, M.: DualGAN: Unsupervised dual learning for image-to-image translation. In: ICCV (2017)
41. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: CVPR (2014)
42. 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. In: ICCV (2017)
43. Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network. In: CVPR (2017)
44. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: ECCV (2016)
45. Zhu, J.Y., Krähenbühl, P., Shechtman, E., Efros, A.A.: Generative visual manipulation on the natural image manifold. In: ECCV (2016)
46. Zhu, J., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)