DualGAN: Unsupervised Dual Learning for Image-to-Image Translation

Zili Yi[1,2], Hao Zhang[2], Ping Tan[2], and Minglun Gong[1]

[1]Memorial University of Newfoundland, Canada
[2]Simon Fraser University, Canada

Abstract

Conditional Generative Adversarial Networks (GANs) for cross-domain image-to-image translation have made much progress recently [7, 8, 21, 12, 4, 18]. Depending on the task complexity, thousands to millions of labeled image pairs are needed to train a conditional GAN. However, human labeling is expensive, even impractical, and large quantities of data may not always be available. Inspired by dual learning from natural language translation [23], we develop a novel dual-GAN mechanism, which enables image translators to be trained from two sets of unlabeled images from two domains. In our architecture, the primal GAN learns to translate images from domain U to those in domain V, while the dual GAN learns to invert the task. The closed loop made by the primal and dual tasks allows images from either domain to be translated and then reconstructed. Hence a loss function that accounts for the reconstruction error of images can be used to train the translators. Experiments on multiple image translation tasks with unlabeled data show considerable performance gain of DualGAN over a single GAN. For some tasks, DualGAN can even achieve comparable or slightly better results than conditional GAN trained on fully labeled data.

1. Introduction

Many image processing and computer vision tasks, e.g., image segmentation, stylization, and abstraction, can be posed as image-to-image translation problems [4], which convert one visual representation of an object or scene into another. Conventionally, these tasks have been tackled separately due to their intrinsic disparities [7, 8, 21, 12, 4, 18]. It is not until the past two years that general-purpose and end-to-end deep learning frameworks, most notably those utilizing fully convolutional networks (FCNs) [11] and conditional generative adversarial nets (cGANs) [4], have been developed to enable a *unified* treatment of these tasks.

Up to date, these general-purpose methods have all been supervised and trained with a large number of *labeled* and *matching* image *pairs*. In practice however, acquiring such training data can be time-consuming (e.g., with pixelwise or patchwise labeling) and even unrealistic. For example, while there are plenty of photos or sketches available, photo-sketch image pairs depicting the same people under the same pose are scarce. In other image translation settings, e.g., converting daylight scenes to night scenes, even though labeled and matching image pairs can be obtained with stationary cameras, moving objects in the scene often cause varying degrees of content discrepancies.

In this paper, we aim to develop an *unsupervised* learning framework for general-purpose image-to-image translation, which only relies on *unlabeled* image data, such as two sets of photos and sketches for the photo-to-sketch conversion task. The obvious technical challenge is how to train a translator without any data characterizing correct translations. Our approach is inspired by *dual learning* from natural language processing [23]. Dual learning trains *two* "opposite" language translators (e.g., English-to-French and French-to-English) simultaneously by minimizing the *reconstruction loss* resulting from a *nested* application of the two translators. The two translators represent a primal-dual pair and the nested application forms a closed loop, allowing the application of reinforcement learning. Specifically, the reconstruction loss measured over monolingual data (either English or French) would generate informative feedback to train a bilingual translation model.

Our work develops a dual learning framework for image-to-image translation for the first time and differs from the original NLP dual learning method of Xia et al. [23] in two main aspects. First, the NLP method relied on pre-trained (English and French) language models to indicate how confident the the translator outputs are natural sentences in their respective target languages. With general-purpose processing in mind and the realization that such pre-trained models are difficult to obtain for many image translation tasks, our work develops GAN discriminators [3] that are trained ad-

versarially with the translators to capture domain distributions. Hence, we call our learning architecture *DualGAN*. Furthermore, we employ FCNs as translators which naturally accommodate the 2D structure of images, rather than sequence-to-sequence translation models such as LSTM or Gated Recurrent Unit (GUT).

Taking two sets of unlabeled images as input, each characterizing an image domain, DualGAN simultaneously learns two reliable image translators from one domain to the other and hence can operate on a wide variety of image-to-image translation tasks. The effectiveness of DuanGAN is validated through comparison with both GAN (with an image-conditional generator and the original discriminator) and conditional GAN [4]. The comparison results demonstrate that, for some applications, DualGAN can outperform supervised methods trained on labeled data.

2. Related work

Since the seminal work by Goodfellow et al. [3] in 2014, a series of GAN-family methods have been proposed for a wide variety of problems. The original GAN can learn a generator to capture the distribution of real data by introducing an adversarial discriminator that evolves to discriminate between the real data and the fake [3]. Soon after, various conditional GANs (cGAN) have been proposed to condition the image generation on class labels [13], attributes [14, 24], texts [15], and images [7, 8, 21, 12, 4, 18].

Most image-conditional models were developed for specific applications such as super-resolution [7], texture synthesis [8], style transfer from normal maps to images [21], and video prediction [12], whereas few others were aiming for general-purpose processing [4, 18]. The general-purpose solution for image-to-image translation proposed by Isola et al. [4] requires significant number of labeled image pairs. The unsupervised mechanism for cross-domain image conversion presented by Taigman et al. [18] can train an image-conditional generator without paired images, but relies on a sophisticated pre-trained function that maps images from either domain to an intermediate representation, which requires labeled data in other formats.

Dual learning was first proposed by Xia et al. [23] to reduce the requirement on labeled data in training English-to-French and French-to-English translators. The French-to-English translation is the dual task to English-to-French translation, and they can be trained side-by-side. The key idea of dual learning is to set up a dual-learning game which involves two agents, each of whom only understands one language, and can evaluate how likely the translated are natural sentences in targeted language and to what extent the reconstructed are consistent with the original. Such a mechanism is played alternatively on both sides, allowing translators to be trained from monolingual data only.

Despite of a lack of parallel bilingual data, two types of feedback signals can be generated: the membership score which evaluates the likelihood of the translated texts belonging to the targeted language, and the reconstruction error that measures the disparity between the reconstructed sentences and the original. Both signals are assessed with the assistance of application-specific domain knowledge, i.e., the pre-trained English and French language models.

In our work, we aim for a general-purpose solution for image-to-image conversion and hence do not utilize any domain-specific knowledge or pre-trained domain representations. Instead, we use a domain-adaptive GAN discriminator to evaluate the membership score of translated samples, whereas the reconstruction error is measured as the mean of absolute difference between the reconstructed and original images within each image domain.

In CycleGAN, a concurrent work by Zhu et al. [26], the same idea for unpaired image-to-image translation is proposed, where the primal-dual relation in DualGAN is referred to as a cyclic mapping and their *cycle consistency loss* is essentially the same as our reconstruction loss. Superiority of CycleGAN has been demonstrated on several tasks where paired training data hardly exist, e.g., in object transfiguration and painting style and season transfer.

Recent work by Liu and Tuzel [10], which we refer to as coupled GAN or CoGAN, also trains two GANs together to solve image translation problems without paired training data. Unlike DualGAN or CycleGAN, the two GANs in CoGAN are not linked to enforce cycle consistency. Instead, CoGAN learns a joint distribution over images from two domains. By sharing weight parameters corresponding to high level semantics in both generative and discriminative networks, CoGAN can enforce the two GANs to interpret these image semantics in the same way. However, the weight-sharing assumption in CoGAN and similar approaches, e.g., [2, 9], does not lead to effective general-purpose solutions as its applicability is task-dependent, leading to unnatural image translation results, as shown in comparative studies by CycleGAN [26].

DualGAN and CycleGAN both aim for general-purpose image-to-image translations without requiring a joint representation to bridge the two image domains. In addition, DualGAN trains both primal and dual GANs at the same time, allowing a reconstruction error term to be used to generate informative feedback signals.

3. Method

Given two sets of unlabeled and unpaired images sampled from domains U and V, respectively, the primal task of DualGAN is to learn a generator $G_A : U \to V$ that maps an image $u \in U$ to an image $v \in V$, while the dual task is to train an inverse generator $G_B : V \to U$. To realize this, we employ two GANs, the primal GAN and the dual GAN. The primal GAN learns the generator G_A and a discrimi-

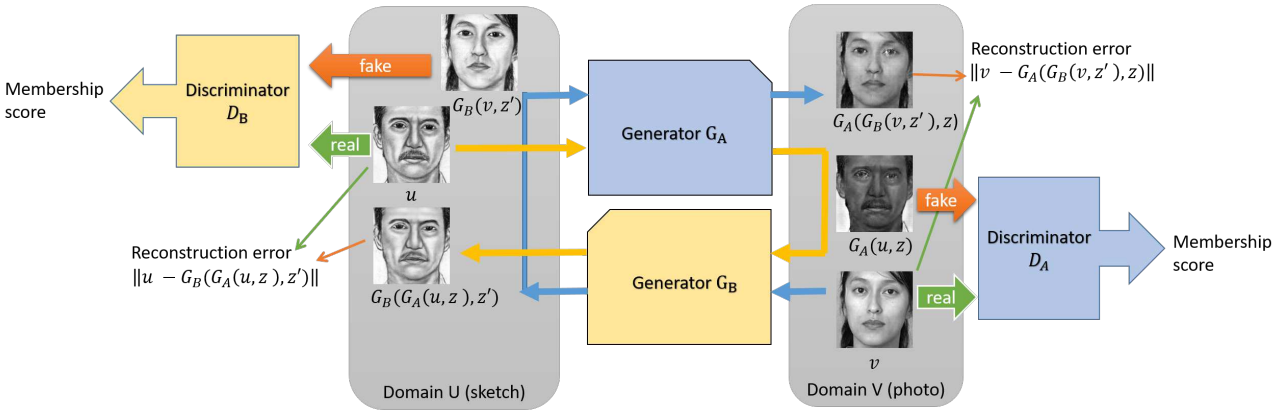

Figure 1: Network architecture and data flow chart of DualGAN for image-to-image translation.

nator D_A that discriminates between G_A's fake outputs and real members of domain V. Analogously, the dual GAN learns the generator G_B and a discriminator D_B. The overall architecture and data flow are illustrated in Fig. 1.

As shown in Fig. 1, image $u \in U$ is translated to domain V using G_A. How well the translation $G_A(u,z)$ fits in V is evaluated by D_A, where z is random noise, and so is z' that appears below. $G_A(u,z)$ is then translated back to domain U using G_B, which outputs $G_B(G_A(u,z),z')$ as the reconstructed version of u. Similarly, $v \in V$ is translated to U as $G_B(v,z')$ and then reconstructed as $G_A(G_B(v,z'),z)$. The discriminator D_A is trained with v as positive samples and $G_A(u,z)$ as negative examples, whereas D_B takes u as positive and $G_B(v,z')$ as negative. Generators G_A and G_B are optimized to emulate "fake" outputs to blind the corresponding discriminators D_A and D_B, as well as to minimize the two *reconstruction losses* $\|G_A(G_B(v,z'),z)-v\|$ and $\|G_B(G_A(u,z),z')-u\|$.

3.1. Objective

As in the traditional GAN, the objective of discriminators is to discriminate the generated fake samples from the real ones. Nevertheless, here we use the loss format advocated by Wasserstein GAN (WGAN) [1] rather than the sigmoid cross-entropy loss used in the original GAN [3]. It is proven that the former performs better in terms of generator convergence and sample quality, as well as in improving the stability of the optimization [1]. The corresponding loss functions used in D_A and D_B are defined as:

$$l_A^d(u,v) = D_A(G_A(u,z)) - D_A(v), \quad (1)$$
$$l_B^d(u,v) = D_B(G_B(v,z')) - D_B(u), \quad (2)$$

where $u \in U$ and $v \in V$.

The same loss function is used for both generators G_A and G_B as they share the same objective. Previous works on conditional image synthesis found it beneficial to replace L_2 distance with L_1, since the former often leads to blurriness [6, 23]. Hence, we adopt L_1 distance to measure the recovery error, which is added to the GAN objective to force the translated samples to obey the domain distribution:

$$\begin{aligned} l^g(u,v) = \lambda_U \|u - G_B(G_A(u,z),z')\| + \\ \lambda_V \|v - G_A(G_B(v,z'),z)\| \\ -D_A(G_B(v,z')) - D_B(G_A(u,z)), \end{aligned} \quad (3)$$

where $u \in U$, $v \in V$, and λ_U, λ_V are two constant parameters. Depending on the application, λ_U and λ_V are typically set to a value within $[100.0, 1,000.0]$. If U contains natural images and V does not (e.g., aerial photo-maps), we find it more effective to use smaller λ_U than λ_V.

3.2. Network configuration

DualGAN is constructed with identical network architecture for G_A and G_B. The generator is configured with equal number of downsampling (pooling) and upsampling layers. In addition, we configure the generator with skip connections between mirrored downsampling and upsampling layers as in [16, 4], making it a U-shaped net. Such a design enables low-level information to be shared between input and output, which is beneficial since many image translation problems implicitly assume alignment between image structures in the input and output (e.g., object shapes, textures, clutter, etc.). Without the skip layers, information from all levels has to pass through the bottleneck, typically causing significant loss of high-frequency information. Furthermore, similar to [4], we did not explicitly provide the noise vectors z, z'. Instead, they are provided only in the form of dropout and applied to several layers of our generators at both training and test phases.

For discriminators, we employ the Markovian PatchGAN architecture as explored in [8], which assumes independence between pixels distanced beyond a specific patch

size and models images only at the patch level rather than over the full image. Such a configuration is effective in capturing local high-frequency features such as texture and style, but less so in modeling global distributions. It fulfills our needs well, since the recovery loss encourages preservation of global and low-frequency information and the discriminators are designated to capture local high-frequency information. The effectiveness of this configuration has been verified on various translation tasks [23]. Similar to [23], we run this discriminator convolutionally across the image, averaging all responses to provide the ultimate output. An extra advantage of such a scheme is that it requires fewer parameters, runs faster, and has no constraints over the size of the input image. The patch size at which the discriminator operates is fixed at 70×70, and the image resolutions were mostly 256×256, same as pix2pix [4].

3.3. Training procedure

To optimize the DualGAN networks, we follow the training procedure proposed in WGAN [1]; see Alg. 1. We train the discriminators n_{critic} steps, then one step on generators. We employ mini-batch Stochastic Gradient Descent and apply the RMSProp solver, as momentum based methods such as Adam would occasionally cause instability [1], and RMSProp is known to perform well even on highly non-stationary problems [19, 1]. We typically set the number of critic iterations per generator iteration n_{critic} to 2-4 and assign batch size to 1-4, without noticeable differences on effectiveness in the experiments. The clipping parameter c is normally set in $[0.01, 0.1]$, varying by application.

Algorithm 1 DualGAN training procedure

Require: Image set U, image set V, GAN A with generator parameters θ_A and discriminator parameters ω_A, GAN B with generator parameters θ_B and discriminator parameters ω_B, clipping parameter c, batch size m, and n_{critic}

Randomly initialize ω_i, θ_i, $i \in \{A, B\}$
repeat
  for $t = 1, \ldots, n_{critic}$ **do**
    sample images $\{u^{(k)}\}_{k=1}^{m} \subseteq U$, $\{v^{(k)}\}_{k=1}^{m} \subseteq V$
    update ω_A to minimize $\frac{1}{m}\sum_{k=1}^{m} l_A^d(u^{(k)}, v^{(k)})$
    update ω_B to minimize $\frac{1}{m}\sum_{k=1}^{m} l_B^d(u^{(k)}, v^{(k)})$
    $clip(\omega_A, -c, c)$, $clip(\omega_B, -c, c)$
  end for
  sample images $\{u^{(k)}\}_{k=1}^{m} \subseteq U$, $\{v^{(k)}\}_{k=1}^{m} \subseteq V$
  update θ_A, θ_B to minimize $\frac{1}{m}\sum_{k=1}^{m} l^g(u^{(k)}, v^{(k)})$
until convergence

Training for traditional GANs needs to carefully balance between the generator and the discriminator, since, as the discriminator improves, the sigmoid cross-entropy loss is locally saturated and may lead to vanishing gradients. Unlike in traditional GANs, the Wasserstein loss is differentiable almost everywhere, resulting in a better discriminator. At each iteration, the generators are not trained until the discriminators have been trained for n_{critic} steps. Such a procedure enables the discriminators to provide more reliable gradient information [1].

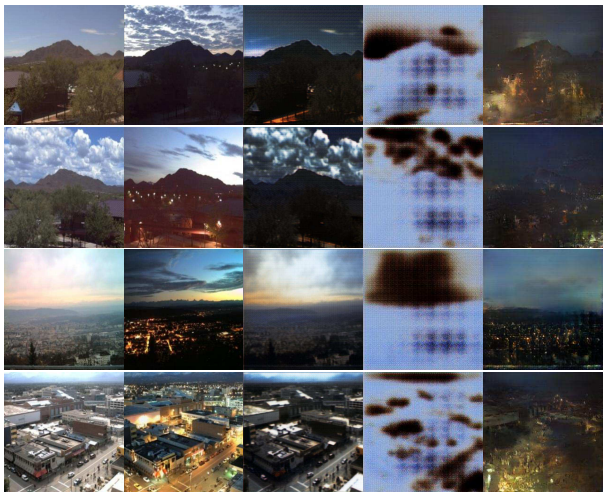

Figure 2: Results of day→night translation. cGAN [4] is trained with labeled data, whereas DualGAN and GAN are trained in an unsupervised manner. DualGAN successfully emulates the night scenes while preserving textures in the inputs, e.g., see differences over the cloud regions between our results and the ground truth (GT). In comparison, results of cGAN and GAN contain much less details.

4. Experimental results and evaluation

To assess the capability of DualGAN in general-purpose image-to-image translation, we conduct experiments on a variety of tasks, including photo-sketch conversion, label-image translation, and artistic stylization.

To compare DualGAN with GAN and cGAN [4], four labeled datasets are used: PHOTO-SKETCH [22, 25], DAY-NIGHT [5], LABEL-FACADES [20], and AERIAL-MAPS, which was directly captured from Google Map [4]. These datasets consist of corresponding images between two domains; they serve as ground truth (GT) and can also be used for supervised learning. However, none of these datasets could guarantee accurate feature alignment at the pixel level. For example, the sketches in SKETCH-PHOTO dataset were drawn by artists and do not accurately align with the corresponding photos, moving objects and cloud pattern changes often show up in the DAY-NIGHT dataset, and the labels in LABEL-FACADES dataset are not always

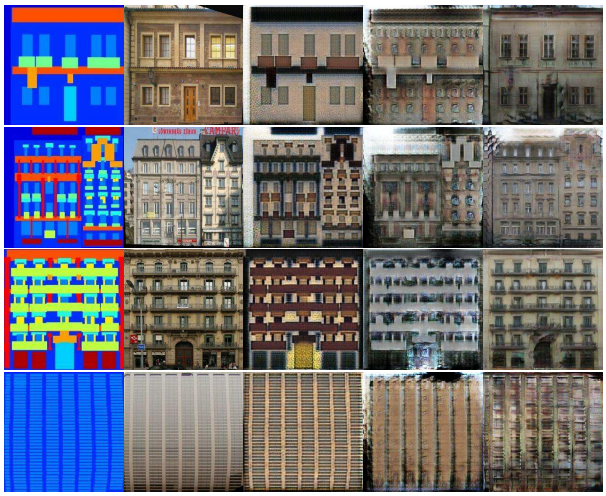

Figure 3: Results of label→facade translation. DualGAN faithfully preserves the structures in the label images, even though some labels do not match well with the corresponding photos in finer details. In contrast, results from GAN and cGAN contain many artifacts. Over regions with label-photo misalignment, cGAN often yields blurry output (e.g., the roof in second row and the entrance in third row).

precise. This highlights, in part, the difficulty in obtaining high quality matching image pairs.

DualGAN enables us to utilize abundant unlabeled image sources from the Web. Two unlabeled and unpaired datasets are also tested in our experiments. The MATERIAL dataset includes images of objects made of different materials, e.g., stone, metal, plastic, fabric, and wood. These images were manually selected from Flickr and cover a variety of illumination conditions, compositions, color, texture, and material sub-types [17]. This dataset was initially used for material recognition, but is applied here for material transfer. The OIL-CHINESE painting dataset includes artistic paintings of two disparate styles: oil and Chinese. All images were crawled from search engines and they contain images with varying quality, format, and size. We reformat, crop, and resize the images for training and evaluation. In both of these datasets, no correspondence is available between images from different domains.

5. Qualitative evaluation

Using the four labeled datasets, we first compare DualGAN with GAN and cGAN [4] on the following translation tasks: day→night (Figure 2), labels↔facade (Figures 3 and 10), face photo↔sketch (Figures 4 and 5), and map↔aerial photo (Figures 8 and 9). In all these tasks, cGAN was trained with labeled (i.e., paired) data, where we ran the model and code provided in [4] and chose the optimal loss function for each task: L_1 loss for facade→label and $L_1 + cGAN$ loss for the other tasks (see [4] for more details). In contrast, DualGAN and GAN were trained in an unsupervised way, i.e., we decouple the image pairs and then reshuffle the data. The results of GAN were generated using our approach by setting $\lambda_U = \lambda_V = 0.0$ in eq. (3), noting that this GAN is different from the original GAN model [3] as it employs a conditional generator.

All three models were trained on the same training datasets and tested on novel data that does not overlap those for training. All the training were carried out on a single GeForce GTX Titan X GPU. At test time, all models ran in well under a second on this GPU.

Compared to GAN, in almost all cases, DualGAN produces results that are less blurry, contain fewer artifacts, and better preserve content structures in the inputs and capture features (e.g., texture, color, and/or style) of the target domain. We attribute the improvements to the reconstruction loss, which forces the inputs to be reconstructable from outputs through the dual generator and strengthens feedback signals that encodes the targeted distribution.

In many cases, DualGAN also compares favorably over the supervised cGAN in terms of sharpness of the outputs and faithfulness to the input images; see Figures 2, 3, 4, 5, and 8. This is encouraging since the supervision in cGAN does utilize additional image and pixel correspondences. On the other hand, when translating between photos and semantic-based labels, such as map↔aerial and label↔facades, it is often impossible to infer the correspondences between pixel colors and labels based on targeted distribution alone. As a result, DualGAN may map pixels to wrong labels (see Figures 9 and 10) or labels to wrong colors/textures (see Figures 3 and 8).

Figures 6 and 7 show image translation results obtained using the two unlabeled datasets, including oil↔Chinese, plastic→metal, metal→stone, leather→fabric, as well as wood↔plastic. The results demonstrate that visually convincing images can be generated by DualGAN when no corresponding images can be found in the target domains. As well, the DualGAN results generally contain less artifacts than those from GAN.

5.1. Quantitative evaluation

To quantitatively evaluate DualGAN, we set up two user studies through Amazon Mechanical Turk (AMT). The "material perceptual" test evaluates the material transfer results, in which we mix the outputs from all material transfer tasks and let the Turkers choose the best match based on which material they believe the objects in the image are made of. For a total of 176 output images, each was evaluated by ten Turkers. An output image is rated as a success if at least three Turkers selected the target material type. Suc-

Figure 4: Photo→sketch translation for faces. Results of DualGAN are generally sharper than those from cGAN, even though the former was trained using unpaired data, whereas the latter makes use of image correspondence.

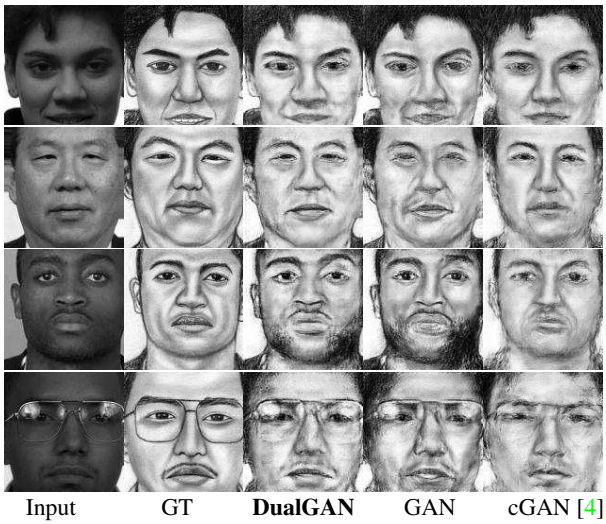

Figure 5: Results for sketch→photo translation of faces. More artifacts and blurriness are showing up in results generated by GAN and cGAN than DualGAN.

cess rates of various material transfer results using different approaches are summarized in Table 1, showing that DualGAN outperforms GAN by a large margin.

In addition, we run the AMT "realness score" evaluation for sketch→photo, label map→facades, maps→aerial photo, and day→night translations. To eliminate potential bias, for each of the four evaluations, we randomly shuffle real photos and outputs from all three approaches before showing them to Turkers. Each image is shown to 20 Turkers, who were asked to score the image based on to what extent the synthesized photo looks real. The "realness" score ranges from 0 (totally missing), 1 (bad), 2 (acceptable), 3 (good), to 4 (compelling). The average score of different approaches on various tasks are then computed and shown in Table. 2. The AMT study results show that DualGAN outperforms GAN on all tasks and outperforms cGAN on two tasks as well. This indicates that cGAN has little tolerance to misalignment and inconsistency between image pairs, but the additional pixel-level correspondence does help cGAN correctly map labels to colors and textures.

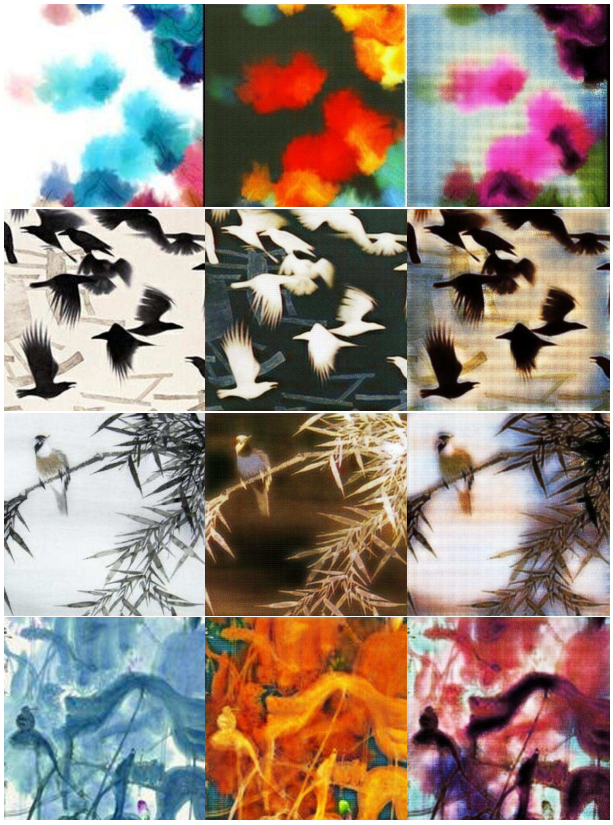

Figure 6: Experimental results for translating Chinese paintings to oil paintings (without GT available). The background grids shown in the GAN results imply that the outputs of GAN are not as stable as those of DualGAN.

Finally, we compute the segmentation accuracies for facades→label and aerial→map tasks, as reported in Tables 3 and 4. The comparison shows that DualGAN is outperformed by cGAN, which is expected as it is difficult to infer proper labeling without image correspondence information from the training data.

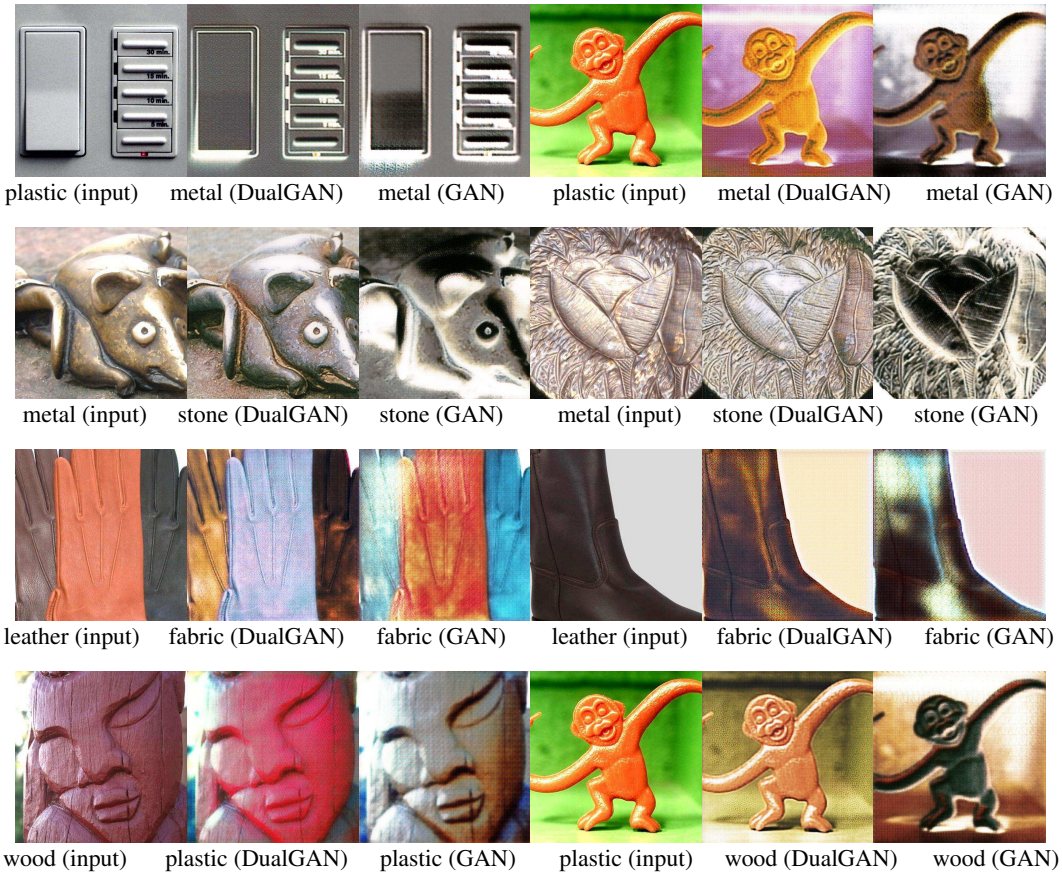

Figure 7: Experimental results for various material transfer tasks. From top to bottom, plastic→metal, metal→stone, leather→fabric, and plastic↔wood.

Task	DualGAN	GAN
plastic→wood	**2**/11	0/11
wood→plastic	**1**/11	0/11
metal→stone	**2**/11	0/11
stone→metal	**2**/11	0/11
leather→fabric	**3**/11	2/11
fabric→leather	**2**/11	1/11
plastic→metal	**7**/11	3/11
metal→plastic	**1**/11	0/11

Table 1: Success rates of various material transfer tasks based on the AMT "material perceptual" test. There are 11 images in each set of transfer result, with noticeable improvements of DualGAN over GAN.

Task	Avg. "realness" score			
	DualGAN	cGAN[4]	GAN	GT
sketch→photo	**1.87**	1.69	1.04	3.56
day→night	**2.42**	1.89	0.13	3.05
label→facades	1.89	**2.59**	1.43	3.33
map→aerial	2.52	**2.92**	1.88	3.21

Table 2: Average AMT "realness" scores of outputs from various tasks. The results show that DualGAN outperforms GAN in all tasks. It also outperforms cGAN for sketch→photo and day→night tasks, but still lag behind for label→facade and map→aerial tasks. In the latter two tasks, the additional image correspondence in training data would help cGAN map labels to the proper colors/textures.

6. Conclusion

We propose DualGAN, a novel unsupervised dual learning framework for general-purpose image-to-image translation. The unsupervised characteristic of DualGAN enables many real world applications, as demonstrated in this work, as well as in the concurrent work CycleGAN [26].

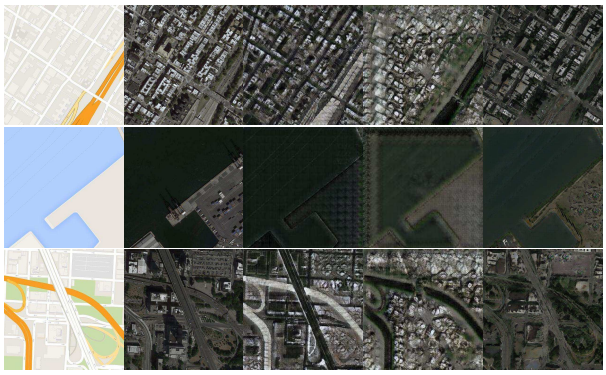

Figure 8: Map→aerial photo translation. Without image correspondences for training, DualGAN may map the orange-colored interstate highways to building roofs with bright colors. Nevertheless, the DualGAN results are sharper than those from GAN and cGAN.

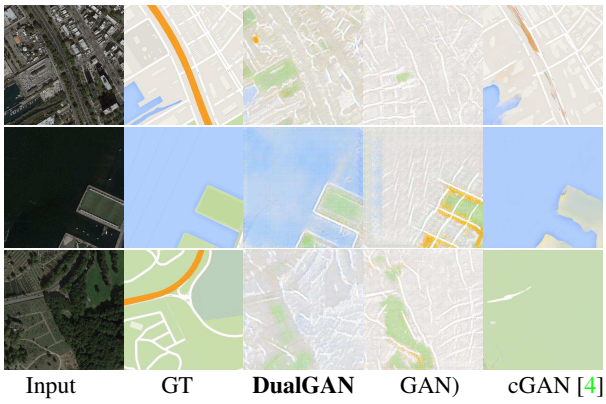

Figure 9: Results for aerial photo→map translation. DualGAN performs better than GAN, but not as good as cGAN. With additional pixel correspondence information, cGAN performs well in terms of labeling local roads, but still cannot detect interstate highways.

Experimental results suggest that the DualGAN mechanism can significantly improve the outputs of GAN for various image-to-image translation tasks. With unlabeled data only, DualGAN can generate comparable or even better outputs than conditional GAN [4] which is trained with labeled data providing image and pixel-level correspondences.

On the other hand, our method is outperformed by conditional GAN or cGAN [4] for certain tasks which involve semantics-based labels. This is due to the lack of pixel and label correspondence information, which cannot be inferred from the target distribution alone. In the future, we intend to investigate whether this limitation can be lifted with the use of a small number of labeled data as a warm start.

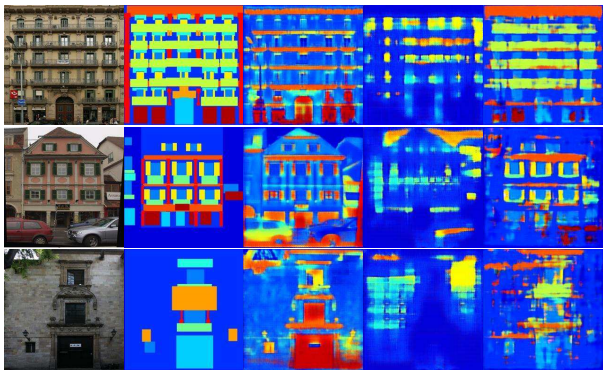

Figure 10: Facades→label translation. While cGAN correctly labels various bulding components such as windows, doors, and balconies, the overall label images are not as detailed and structured as DualGAN's outputs.

	Per-pixel acc.	Per-class acc.	Class IOU
DualGAN	0.27	0.13	0.06
cGAN [4]	**0.54**	**0.33**	**0.19**
GAN	0.22	0.10	0.05

Table 3: Segmentation accuracy for the facades→label task. DualGAN outperforms GAN, but is not as accurate as cGAN. Without image correspondence (for cGAN), even if DualGAN segments a region properly, it may not assign the region with a correct label.

	Per-pixel acc.	Per-class acc.	Class IOU
DualGAN	0.42	0.22	0.09
cGAN [4]	**0.70**	**0.46**	**0.26**
GAN	0.41	0.23	0.09

Table 4: Segmentation accuracy for the aerial→map task, for which DualGAN performs less than satisfactorily.

Acknowledgment. We thank all the anonymous reviewers for their valuable comments and suggestions. The first author is a PhD student from the Memorial University of Newfoundland and has been visiting SFU since 2016. This work was supported in part by grants from the Natural Sciences and Engineering Research Council (NSERC) of Canada (No. 611370, 2017-06086).

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3, 4

[2] Y. Aytar, L. Castrejon, C. Vondrick, H. Pirsiavash, and A. Torralba. Cross-modal scene networks. *CoRR*, abs/1610.09003, 2016. 2

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 2, 3, 5

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[5] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM Transactions on Graphics (TOG)*, 33(4):149, 2014. 4

[6] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015. 3

[7] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 1, 2

[8] C. Li and M. Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 702–716. Springer, 2016. 1, 2, 3

[9] M. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. 2

[10] M.-Y. Liu and O. Tuzel. Coupled generative adversarial networks. In *Advances in neural information processing systems*, pages 469–477, 2016. 2

[11] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. 1

[12] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 1, 2

[13] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[14] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 2

[15] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 3, 2016. 2

[16] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015. 3

[17] L. Sharan, R. Rosenholtz, and E. Adelson. Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8):784–784, 2009. 5

[18] Y. Taigman, A. Polyak, and L. Wolf. Unsupervised cross-domain image generation. *arXiv preprint arXiv:1611.02200*, 2016. 1, 2

[19] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012. 4

[20] R. Tyleček and R. Šára. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, pages 364–374. Springer, 2013. 4

[21] X. Wang and A. Gupta. Generative image modeling using style and structure adversarial networks. In *European Conference on Computer Vision (ECCV)*, pages 318–335. Springer, 2016. 1, 2

[22] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2009. 4

[23] Y. Xia, D. He, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma. Dual learning for machine translation. *arXiv preprint arXiv:1611.00179*, 2016. 1, 2, 3, 4

[24] X. Yan, J. Yang, K. Sohn, and H. Lee. Attribute2image: Conditional image generation from visual attributes. In *European Conference on Computer Vision (ECCV)*, pages 776–791. Springer, 2016. 2

[25] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 513–520. IEEE, 2011. 4

[26] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV), to appear*, 2017. 2, 7