Deliberation Learning for Image-to-Image Translation

Tianyu He[1] *, **Yingce Xia**[2] *, **Jianxin Lin**[1], **Xu Tan**[2], **Di He**[3], **Tao Qin**[2] and **Zhibo Chen**[1] †
[1]CAS Key Laboratory of Technology in Geo-spatial Information Processing and Application System,
University of Science and Technology of China
[2]Microsoft Research Asia
[3]Key Laboratory of Machine Perception, MOE, School of EECS, Peking University
{hetianyu, linjx}@mail.ustc.edu.cn, {Yingce.Xia, xuta}@microsoft.com, di_he@pku.edu.cn,
taoqin@microsoft.com, chenzhibo@ustc.edu.cn

Abstract

Image-to-image translation, which transfers an image from a source domain to a target one, has attracted much attention in both academia and industry. The major approach is to adopt an encoder-decoder based framework, where the encoder extracts features from the input image and then the decoder decodes the features and generates an image in the target domain as the output. In this paper, we go beyond this learning framework by considering an additional polishing step on the output image. Polishing an image is very common in human's daily life, such as editing and beautifying a photo in Photoshop after taking/generating it by a digital camera. Such a deliberation process is shown to be very helpful and important in practice and thus we believe it will also be helpful for image translation. Inspired by the success of deliberation network in natural language processing, we extend deliberation process to the field of image translation. We verify our proposed method on four two-domain translation tasks and one multi-domain translation task. Both the qualitative and quantitative results demonstrate the effectiveness of our method.

1 Introduction

Unsupervised image-to-image translation is an important application task in computer vision [Zhu *et al.*, 2017; Choi *et al.*, 2018]. The encoder-decoder framework is widely used to achieve such translation, where the encoder maps the image to a latent representation and the decoder translates it to the target domain.

Considering labeled data is costly to obtain, unsupervised image-to-image translation is widely adopted, which tries to uncover the mapping without paired images in two domains. CycleGAN [Zhu *et al.*, 2017] is one of the most cited model that can achieve unsupervised translation between two domains. Denote $\mathcal{X}_s$ and $\mathcal{X}_t$ as two image domains of interests, and no labeled image pair in $\mathcal{X}_s \times \mathcal{X}_t$ is available. As the name suggests, the two ingredients to make CycleGAN successful are: (1) Minimizing the reconstruction loss between two models $\mathcal{X}_s \to \mathcal{X}_t \to \mathcal{X}_s$ and $\mathcal{X}_t \to \mathcal{X}_s \to \mathcal{X}_t$; such a loss can ensure the main content of an image is kept; (2) the adversarial network [Goodfellow *et al.*, 2014], that can verify whether an output image belongs to specific domain. DualGAN [Yi *et al.*, 2017] and DiscoGAN [Kim *et al.*, 2017] also adopt similar ideas. Another important model StarGAN [Choi *et al.*, 2018] is designed for multi-domain image-to-image translation with one encoder and one decoder. The motivation of StarGAN is to reduce parameters by using a single model to achieve multi-domain translation instead of using multiple independent models.

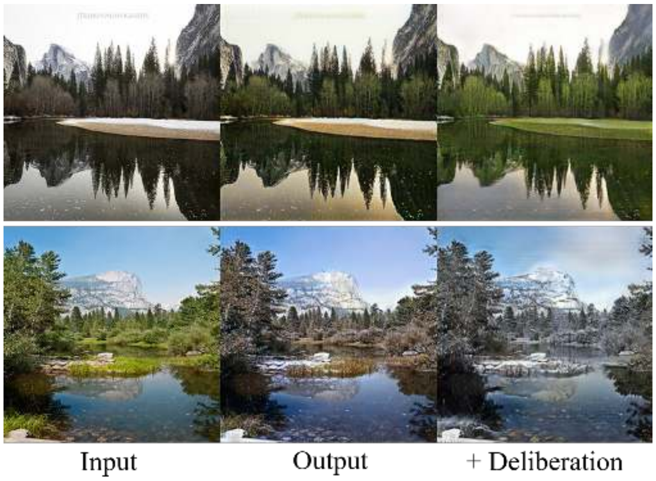

Figure 1: Examples of Winter→Summer translation (top row) and Summer→Winter translation (bottom row), where the three columns are the input, output of standard CycleGAN and output after deliberation.

Although the aforementioned techniques achieved great success, they lack an obvious step compared with human behaviors: deliberation, which means reviewing and keeping polishing the output. For example, to draw a paint, the artist first sketches the outline to get an overall impression of the image, and then gradually enrich more details and textures. Such a behavior broadly exists in human's behavior, but still lacks in unsupervised image-to-image translation literature.

*Equal contribution.

†Corresponding author.

Deliberation is important for image translation as illustrated in Figure 1. For the Winter→Summer translation (first row), there are still some snow in the direct output of the decoder (middle). Similarly, for Summer→Winter translation (second row), we can still find green trees in the output (middle). This shows the necessity of deliberation in image translation. To leverage such an important property, we propose deliberation learning for image-to-image translation, that can further polish and improve the output of a standard model. Take the translation from $\mathcal{X}_s$ to $\mathcal{X}_t$ as an example. Our deliberation learning framework consists of an encoder, a decoder and a post-editor: The encoder and decoder are the same as those in CycleGAN or StarGAN, serving for encoding the image as a hidden vector and decoding the image conditioned on the hidden vector. They work together to translate $x \in \mathcal{X}_s$ to a $\hat{y} \in \mathcal{X}_t$. The post-editor will eventually output another $y^* \in \mathcal{X}_t$, with both x and $\hat{y}$ as inputs. Compared with the standard encoder-decoder framework, the post-editor can have an overall impression of $\hat{y}$, which is the mapping of x in the target domain, and then keep refining on it, while the standard one cannot. As shown in Figure 1, after using post-editor, the snow are erased for Winter→Summer translation and the green parts are enveloped with white frost.

2 Related Work

The generative adversarial networks [Goodfellow *et al.*, 2014] (briefly, GAN) has enabled significant progress in unsupervised image-to-image translation [Zhu *et al.*, 2017; Choi *et al.*, 2018]. There are two essential elements in a GAN: a generator, used to map a random noise to an image; and a discriminator, used to verify whether the input is a natural image or a faked image produced by the generator.

Unsupervised image-to-image translation is an important application of GAN, which means to map an image from one domain to another and has received considerable attention recently. According to the number of domains involved in the translation, the related work can be categorized as two-domain translation and multi-domain translation.

2.1 Two-Domain Translation

Let $\mathcal{X}_s$ and $\mathcal{X}_t$ denote two different image domains. Our target is to learn a mapping $f : \mathcal{X}_s \mapsto \mathcal{X}_t$. A common technique adopted in two-domain translation is the conditional GAN, who have made much progress recently [Mirza and Osindero, 2014; Isola *et al.*, 2017]. In these frameworks, the input image is compressed into a hidden vector via a series of convolutional layers and then converted to the target domain by several transposed convolutional layers. The generated images will be fed into the discriminator to ensure the generation quality. Additional input like random noise [Isola *et al.*, 2017], text [Reed *et al.*, 2016] can also be included.

Inspired by the success of dual learning in neural machine translation [He *et al.*, 2016; Wang *et al.*, 2019], learning two dual mappings $f : \mathcal{X}_s \mapsto \mathcal{X}_t$ and $g : \mathcal{X}_t \mapsto \mathcal{X}_s$ together are introduced [Kim *et al.*, 2017; Zhu *et al.*, 2017; Yi *et al.*, 2017; Lin *et al.*, 2018; Mejjati *et al.*, 2018] to image translation. CycleGAN [Zhu *et al.*, 2017] is one of the most cited work with such an idea: Given an $x \in \mathcal{X}_s$, it is first mapped to $\hat{y}$ by $f(x)$ and then mapped back to $\hat{x}$ by $g(\hat{y})$. The cycle-consistent loss $\|x - \hat{x}\|_1$ is equipped to minimize the distance between x and $\hat{x}$, allowing f and g to obtain feedback signal from its counterpart. Similar idea also exists in DualGAN [Yi *et al.*, 2017] and DiscoGAN [Kim *et al.*, 2017].

2.2 Multi-Domain Translation

Great performances were achieved in dual-domain translation. However, while applying in multi-domain translation, training models for each pair of domains incurs much more resource consumption. To alleviate this limitation of scalability, several works extend dual-domain translation to multi-domain translation by learning relationships among multiple domains in an unsupervised fashion [Choi *et al.*, 2018; Liu *et al.*, 2018; Anoosheh *et al.*, 2018; Pumarola *et al.*, 2018]. Instead of employing a generator and a discriminator for each domain, StarGAN [Choi *et al.*, 2018] learns all mappings f with only one generator and one discriminator.

Although deliberation learning is not widely studied in image generation, it has been used in many natural language processing tasks including neural machine translation [Xia *et al.*, 2017], grammar check [Ge *et al.*, 2018], review generation [Guu *et al.*, 2018]. It is beneficial to introduce this idea into image generation tasks.

3 Framework

In this section, we introduce the framework of *deliberation learning* for image translation, including both two-domain translation (based on CycleGAN) and multi-domain image translation (based on StarGAN).

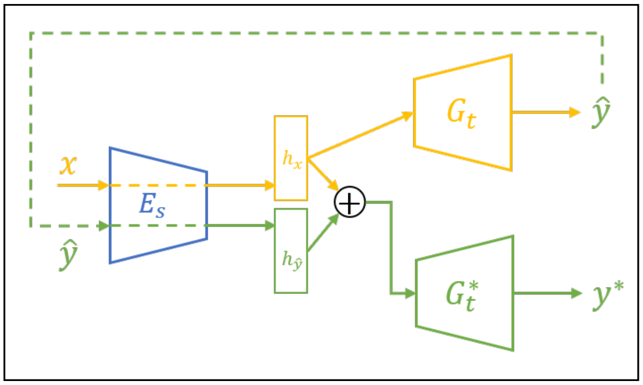

Figure 2: Translation module of $\mathcal{X}_s \to \mathcal{X}_t$, where $x \in \mathcal{X}_s$ and $\hat{y}, y^* \in \mathcal{X}_t$. E_s, G_t and G_t^* denote source domain encoder, the decoder and post-editor in the target domain.

3.1 Two-Domain Translation

Given two image domains $\mathcal{X}_s$ and $\mathcal{X}_t$, the architecture of our proposed deliberation network achieving $\mathcal{X}_s$ to $\mathcal{X}_t$ translation is shown in Figure 2. There are three components: an encoder E_s, a decoder G_t[1], and a post-editor G_t^*. The three

[1] We use $G_\cdot$ instead of $D_\cdot$ to represent the decoder, in order to avoid confusion with the discriminator $D_\cdot$.

components work together to achieve the translation, which is shown in Eqn. (1): for any $x \in \mathcal{X}_s$,

$$\begin{aligned} &h_x = E_s(x);\ \hat{y} = G_t(h_x); \\ &h_{\hat{y}} = E_s(\hat{y});\ y^* = G_t^*(h_x + h_{\hat{y}}). \end{aligned} \tag{1}$$

The y^* is used as the output of the translation. Note that $\hat{y}$ is the output of the conventional models like CycleGAN and DualGAN, without deliberation included. To generate y^*, the information from both raw input x and the first-round output $\hat{y}$ are leveraged, carried by h_x and $h_{\hat{y}}$, and thus leading to better translation results.

The reason we use one encoder E_s to encode both x and $\hat{y}$ is that, in standard CycleGAN, there is an identity mapping loss $\|G_t(E_s(y)) - y\|_1$, $y \in \mathcal{X}_t$. That is, an encoder can naturally encode the images from the target space. To reduce memory cost, we reuse the encoder. We found using two encoders will not bring much additional gain, see Section 4.3.

Following the common practice in image translation, we apply the cycle consistency loss between the translation of two domains as well as the adversarial loss. To achieve $\mathcal{X}_t \to \mathcal{X}_s$ translation, another groups of encoder E_B, decoder G_A and post-editor G_A^* are needed, which work as follows: for any $y \in \mathcal{X}_t$,

$$\begin{aligned} &h_y = E_t(y);\ \hat{x} = G_s(h_y); \\ &h_{\hat{x}} = E_t(\hat{x});\ x^* = G_s^*(h_y + h_{\hat{x}}). \end{aligned} \tag{2}$$

For ease of reference, let f^* denote $\mathcal{X}_s \to \mathcal{X}_t$ translation with Eqn. (1) and g^* denote $\mathcal{X}_t \to \mathcal{X}_s$ translation using Eqn. (2). Note that $G_t \circ E_s$ and $G_s \circ E_t$ can also achieve the above two translations, we $\circ$ denotes the cascade of functions. Denote $G_t \circ E_s$ as f and denote $G_s \circ E_t$ as g respectively.

To stabilize the training, the E_s, E_t, G_s and G_t are pre-trained following standard CycleGAN until convergence. Only G_s^* and G_t^* are updated. An empirical study is shown in Section 4.3. With a little bit confusion, in the next context, $\mathcal{X}_s$ and $\mathcal{X}_t$ also refer to the training datasets of images in the source and target domains.

Cycle consistency loss. Following the common practice in image translation, the cycle consistency loss is defined as

$$\begin{aligned} \ell_{\text{cyc}} =& \frac{1}{|\mathcal{X}_s|}\sum_{x\in\mathcal{X}_s} \|g^*(f^*(x)) - x\|_1 \\ &+ \frac{1}{|\mathcal{X}_t|}\sum_{y\in\mathcal{X}_t} \|f^*(g^*(y)) - y\|_1, \end{aligned} \tag{3}$$

where $|\mathcal{X}_s|$ refers to the number of images in $\mathcal{X}_s$, and so does $|\mathcal{X}_t|$. Both f^* and g^* jointly work to minimize the reconstruction loss, which is exactly what is used standard CycleGAN.

Adversarial loss. To force the generated images belonging to the corresponding domains, we need to use the adversarial loss. Define D_s and D_t as two discriminators of domain $\mathcal{X}_s$ and domain $\mathcal{X}_t$, which can map an input image to $[0, 1]$, indicating the probability that the input is a natural image in the corresponding domain. The adversarial loss is

$$\begin{aligned} \ell_{\text{adv}} =& \frac{1}{2|\mathcal{X}_s|}\sum_{x\in\mathcal{X}_s} (\log D_s(x) + \log(1 - D_t(f^*(x)))) \\ &+ \frac{1}{2|\mathcal{X}_t|}\sum_{y\in\mathcal{X}_t} (\log D_t(y) + \log(1 - D_s(g^*(y)))). \end{aligned} \tag{4}$$

Therefore, the training objective functions for f^* and g^* is to minimize

$$\ell_{\text{total}} = \ell_{\text{cyc}} + \lambda\ell_{\text{adv}}, \tag{5}$$

while the D_s and D_t will work on maximize ℓ_{adv}. In experiments, we fix λ as 10 following [Zhu *et al.*, 2017].

3.2 Multi-Domain Translation

Several works target at image translation among multiple domains [Choi *et al.*, 2018; Liu *et al.*, 2018]. Our proposed deliberation learning framework also works for this setting.

Let $\mathcal{X}_1$, $\mathcal{X}_2$, $\cdots$, $\mathcal{X}_N$ be N domain of interests ($N \geq 2$). Our target is to achieve $N(N-1)$ mappings among the N image domains. It is impractical to learn so many mappings, especially when N is large. An light-weight way is to use a StarGAN like structure, where there is only one encoder, one decoder and one discriminator. Adapted to the deliberation learning framework, there are one encoder E, one decoder G, one post-editor G^* and a discrimiator D. Each image domain has a learnable embedding t_i $i \in [N]$, representing the domain characteristic. The E can map images from any domain to hidden representations conditioned on the embedding; G and G^* can map the hidden representation guided by the domain embedding to the target space. Take the mapping from $\mathcal{X}_i$ to $\mathcal{X}_j$ as an example ($i \neq j$): for an $x \in \mathcal{X}_i$,

$$\begin{aligned} &h_{i\to j} = E(\texttt{concat}[x; t_j]);\ \hat{y} = G(h_{i\to j}); \\ &h_j = E(\texttt{concat}[\hat{y}; t_j]);\ y^* = G^*(h_j + h_{i\to j}), \end{aligned} \tag{6}$$

where `concat` represents padding the second input to the first one along the last dimension. y^* is eventually used as the output of x in $\mathcal{X}_j$. In this case, for ease of reference, denote the translation from $\mathcal{X}_i$ to $\mathcal{X}_j$ following Eqn. (6) as $f_{i,j}^*$. Correspondingly, the generation function based on $G \circ E$ is denoted as $f_{i,j}$. For multi-domain translation, E and G are pre-trained and then fixed too.

Similar to two-domain translation, the training loss of multi-domain translation consists of two parts too: the cycle consistency loss and adversarial loss. The training process of multi-domain deliberation is shown as follows:

1. Randomly choose two different domains $\mathcal{X}_i$ and $\mathcal{X}_j$, where $i \neq j$; randomly sample two batches of data $\mathcal{B}_i \subset \mathcal{X}_i$ and $\mathcal{B}_j \subset \mathcal{X}_j$;
2. Formulate the cycle consistency loss as follows:

$$\begin{aligned} \ell_{\text{cyc}} =& \frac{1}{|\mathcal{B}_i|}\sum_{x\in\mathcal{B}_i} \|f_{j,i}^*(f_{i,j}^*(x)) - x\|_1 \\ &+ \frac{1}{|\mathcal{B}_j|}\sum_{y\in\mathcal{B}_j} \|f_{i,j}^*(f_{j,i}^*(y)) - y\|_1; \end{aligned} \tag{7}$$

3. The discriminator slightly differs from those in two-domain translation. It consists of two parts: D_{src}, which is used to justify whether the input is a natural image; D_{cls}, which is used to verify which domain the image belongs to. D_{src} and D_{cls} share the basic architecture

expect for the last few layers. The adversarial loss is

$$\begin{aligned}\ell_{\text{adv}} =&\frac{1}{2|\mathcal{B}_i|}\sum_{x\in\mathcal{B}_i}(\log D_{\text{src}}(x) + \log(1 - D_{\text{src}}(f^*_{i,j}(x)))\\ &+\frac{1}{2|\mathcal{B}_j|}\sum_{y\in\mathcal{B}_j}(\log D_{\text{src}}(y) + \log(1 - D_{\text{src}}(f^*_{j,i}(y)))\\ &+\frac{1}{2|\mathcal{B}_i|}\sum_{x\in\mathcal{B}_i}(\log D_{\text{cls}}(i|x) + \log D^-_{\text{cls}}(j|f^*_{i,j}(x))\\ &+\frac{1}{2|\mathcal{B}_j|}\sum_{y\in\mathcal{B}_j}(\log D_{\text{cls}}(j|y) + \log D^-_{\text{cls}}(i|f^*_{j,i}(y)).\end{aligned} \tag{8}$$

$f^*_{\cdot,\cdot}$ will work on minimizing ℓ_{adv} while the D_{adv} and D_{cls} will try to enlarge it. When optimizing the discriminators, the D^-_{cls} is fixed.

4. Minimize $\ell_{\text{cyc}} + \lambda\ell_{\text{adv}}$ on $\mathcal{B}_i$ and $\mathcal{B}_j$, where $\lambda = 10$. Repeat step (1) to (4) until convergence.

Compared with two-domain deliberation learning, each component of our model in multi-domain setting has to deal with images from different domains, while that in two-domain setting only works for two-domain.

3.3 Discussion

The idea of deliberation learning also exists in super resolution (briefly, SR), whose task is to convert a low-resolution image to a high-resolution one. A bicubic interpolation [Dong *et al.*, 2014] is first applied to the low-resolution image, then followed by a neural network to reconstruct the high-resolution one. Indeed, the task of SR itself and our framework share the high-level idea, but there are still several differences: (1) Image translation covers at least two domains with different semantics, while SR works on one domain only; (2) When making deliberation, our framework takes the information from two domains as input and further deliberates them by a post-editor. In comparison, SR conducts one-pass operation, where the interpolated image is directly used in the subsequent module.

Another work leveraging deliberation learning is [Xu *et al.*, 2018], which attacks the text-to-image problem: the images are generated from low-resolution to high-resolution, with multi-modal loss as constraints. Different from our work, our model is optimized in an fully unsupervised manner, and what we focused is to polish a generated image (with canonical output resolution) to a better one.

4 Application to Two-Domain Translation

For two-domain translation, we work on four datasets to verify the effectiveness of our algorithm.

4.1 Settings

Tasks. We select four tasks evaluated in CycleGAN [Zhu *et al.*, 2017]: semantic Label↔Photo translation on Cityscapes dataset [Cordts *et al.*, 2016], Apple↔Orange translation, Winter↔Summer translation, and Photo↔Paint translation.

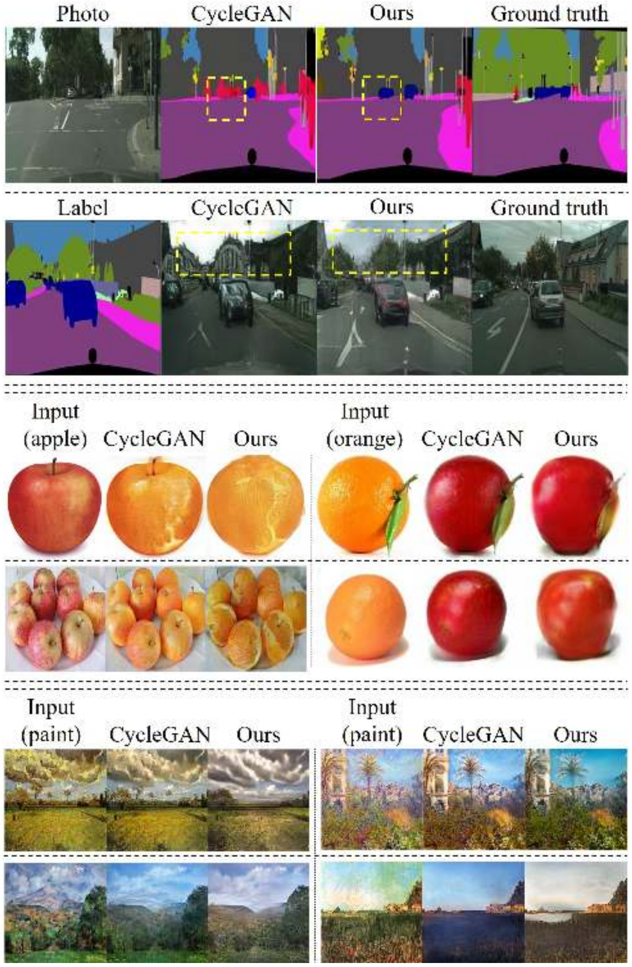

Figure 3: From top to bottom are results of Label↔Photo translation, Apple↔Orange translation, Paint→Photo translation.

Model architecture. For the encoder, decoder and the discriminator, we adopt the same architectures as those in CycleGAN for consistency. In addition, we need two post-editors G^*_s and G^*_t. We split the generator of CycleGAN into two components: The first one serves as the encoder in our scheme, which contains two stride-2 convolutional layers and four residual blocks. The remaining part serves as a decoder, which contains five residual blocks and two $\frac{1}{2}$-strided convolutional layers. Therefore, the total number of layers are consistent with CycleGAN. The architecture of post-editor is the same as the decoder in CycleGAN as we introduced before. For the discriminator, we directly follows CycleGAN.

Implementation details. We follow the offical CycleGAN[2] to implement our scheme in PyTorch. To stabilize the training, We first pre-train the E_s, G_t, E_t and G_s using standard CycleGAN code until convergence. After that, we start to train G^*_s and G^*_t. We use Adam with initial learning rate 2×10^{-4} to train the models for the first 100 epochs. Then we linearly decay the learning rate to 0 in the next 100 epochs.

[2] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix

Task	CycleGAN	Ours
Apple→Orange	147.31	**130.17**
Orange→Apple	146.96	**132.70**
Photo→Label	68.66	**45.62**
Label→Photo	95.23	**69.34**
Winter→Summer	85.48	**76.89**
Summer→Winter	82.20	**77.64**
Cezanne→Photo	186.16	**158.75**
Photo→Cezanne	192.25	**175.78**
Monet→Photo	134.01	**123.96**
Photo→Monet	139.37	**123.45**
Ukiyo-e→Photo	197.46	**162.53**
Photo→Ukiyo-e	152.86	**127.18**
Van Gogh→Photo	93.04	**77.34**
Photo→Van Gogh	96.37	**87.93**

Table 1: FID scores of CycleGAN and our algorithm.

Metrics. Fréchet Inception Distance (briefly, FID) is first proposed by [Heusel *et al.*, 2017], recently becomes a commonly adopted approach to evaluate generative models [Lucic *et al.*, 2018; Brock *et al.*, 2019]. For FID measurement, the generated samples and the real ones are first mapped into feature space by an Inception-v3 model [Szegedy *et al.*, 2016]. Then Fréchet distance between these two distributions is calculated to obtain FID score. The authors demonstrated that FID score has a reasonable correlation with human judgment [Heusel *et al.*, 2017]. Smaller FID scores indicate better translation qualities. In addition, we follow [Zhu *et al.*, 2017] and provide FCN score [Isola *et al.*, 2017][3] in our framework analysis for fair comparison, including per-pixel accuracy (Pixel Acc.), per-class accuracy (Class Acc.) and mean class Intersection-Over-Union (Class IOU). Higher accuracies indicate better translation qualities.

4.2 Results

Qualitative evaluation. Figure 3 shows three groups of translation results on Label↔Photo (the first two rows), Apple↔Orange (the second two rows) and Paint↔Photo (the last two rows). In general, deliberation learning can: (1) Generate images with rich details. (2) Correct many failure cases (e.g.missing items, flawed translation, etc.) generated by CycleGAN. (3) Make the generated images more realistic.

Quantitative evaluation. FID scores of four datasets are shown in Table 1. We can see that deliberation learning significantly surpasses the baseline across all datasets.

4.3 Framework Analysis

We carry out detailed analysis of our proposed framework. We implement another six settings on Label→Photo translation and the scores are listed in Table 2.

(1) To verify whether different ways to combine the features from the encoder and decoder will influence deliberation quality, we concatenate the two features h_x and $h_{\hat{y}}$ in Eqn. (1) instead of adding them. (Ours-1)

Setting	Pixel Acc.	Class Acc.	Class IOU	FID
CoGAN	0.40	0.10	0.06	–
BiGAN	0.19	0.06	0.02	–
CycleGAN	0.52	0.17	0.11	95.23
pix2pix	0.71	0.25	0.18	–
Ours	**0.62**	**0.20**	**0.15**	**69.34**
Ours-1	0.61	0.20	0.15	67.63
Ours-2	0.57	0.19	0.14	81.59
Ours-3	0.57	0.20	0.15	80.83
Ours-4	0.57	0.20	0.15	77.58
Ours-5	0.52	0.13	0.10	106.28
Ours-6	0.56	0.20	0.15	78.29

Table 2: FCN and FID scores of Label→Photo on Cityscapes.

(2) To verify whether better accuracies are brought by larger models, we increase the depth of decoder as double, i.e.ten residual blocks. (Ours-2)

(3) To verify whether we need two different encoders to encode two different outputs, we try to use two encoders to encode x and $\hat{y}$ separately. (Ours-3)

(4) To verify whether we need identity loss, we remove it from our framework. (Ours-4)

(5) To verify whether we need to fix encoder and decoder, we update all parameters in the encoder, decoder and post-editor. (Ours-5)

(6) To verify whether we need to fetch raw input for the post-editor, we only feed the post-editor with $h_{\hat{y}}$. (Ours-6)

For fair comparison, we also provide the FCN scores in Table 2. We list the existing results of CoGAN [Liu and Tuzel, 2016], BiGAN [Donahue *et al.*, 2017; Dumoulin *et al.*, 2017], standard CycleGAN [Zhu *et al.*, 2017] and pix2pix [Isola *et al.*, 2017]. Note that pix2pix is trained in a supervised way.

5 Application to Multi-Domain Translation

In this section, we conduct our deliberation learning on multi-domain image translation. We also give qualitative and quantitative analysis on the performance of our scheme.

5.1 Settings

Tasks. We used the publicly available CelebA dataset [Liu *et al.*, 2015] for facial attributes translation. CelebA contains $10,177$ number of identities and $202,599$ number of facial images. The original images are cropped to 128×156. The test set is randomly sampled ($2,000$ images) and the remaining images are used for training. For directly comparison with StarGAN [Choi *et al.*, 2018], we perform our experiments on the same three attributes: hair color (*black*, *blond* and *brown*), gender (*male* / *female*) and age (*young* / *old*).

Model architecture. For fair comparison, we adopt the same architectures as StarGAN [Choi *et al.*, 2018] for the encoder and decoder. We additionally need one post-editor G^* for our method. Similar to two-domain image translation, we split the generator of StarGAN into encoder and decoder. We adopt the same architecture of decoder for the post-editor G^*. Different from conventional GAN [Goodfellow *et al.*, 2014],

[3] We directly use the code and pre-trained FCN model in https://github.com/phillipi/pix2pix/tree/master/scripts/eval_cityscapes.

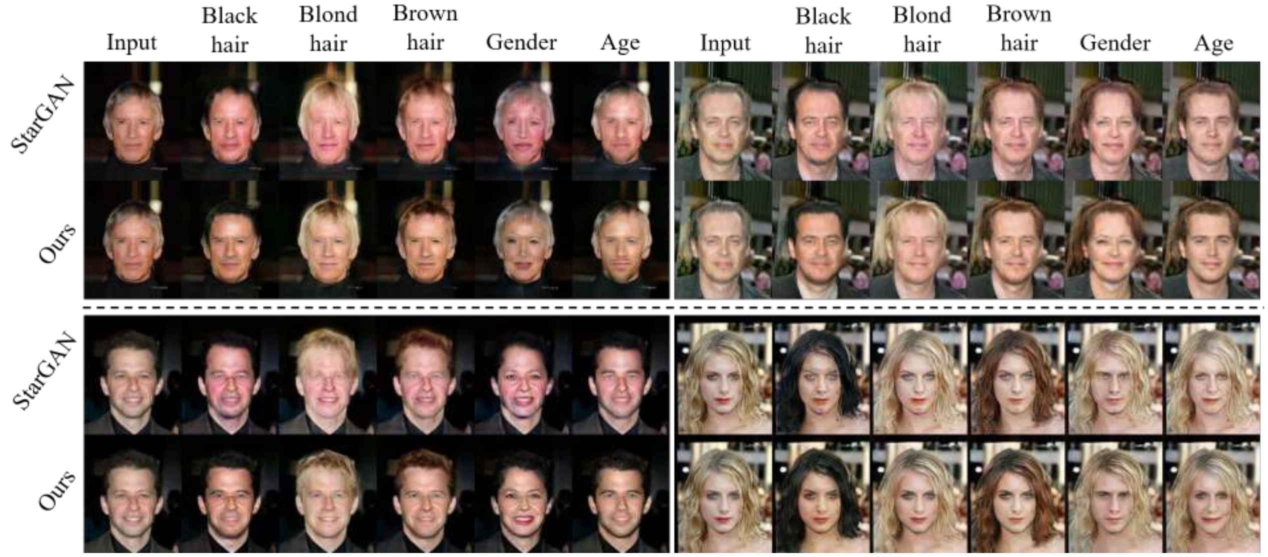

Figure 4: Qualitative results of our scheme compared with StarGAN. It is worth noting that deliberation effectively boost the performance of StarGAN on facial attributes, making the generated images more natural and indistinguishable from the real images.

	Classification Accuracies	
Attributes	StarGAN	Ours
Hair	80.99%	84.10%
Gender	88.40%	94.19%
Age	74.48%	76.64%

Table 3: Classification accuracies of StarGAN and our algorithm.

the discriminator is enhanced with the ability of distinguishing which domain the translated images belong to.

Implementation details. Our implementation based on official code of StarGAN[4]. Similar to CycleGAN, we first pretrain the E_s and G_t with standard StarGAN until convergence. The pre-trained parameters is loaded as initial model to optimize G^*. We use Adam with initial learning rate 1×10^{-4} to train the models for the first $100,000$ iterations. Then we linearly decay the learning rate to 0 in the next $100,000$ iterations. The batch size is set to 16.

Metrics. To evaluate qualities of generated images quantitatively, we follow [Choi *et al.*, 2018] to train a classification model on three facial attributes we used (i.e.hair color, gender and age). We directly adopt the model architecture of discriminator used in StarGAN for the classifier and use the same training set as our image translation models. Higher classification accuracies on the translated facial attributes indicates the better translation qualities.

5.2 Results

Qualitative evaluation. We provide facial attributes translation in Figure 4. Generally, deliberation learning effectively boost the performance of StarGAN on facial attributes, making the generated images more natural and indistinguishable from the real images. From Figure 4, we can observe that, deliberation learning successfully generates facial images with rich details, especially on semantic regions like eye, cheek etc. In addition, in many cases, StarGAN suffers from color shifting, while our scheme correct this fault and preserve the reality without impact on background or hair color.

Quantitative evaluation. The quantitative evaluation of three facial attributes translation are illustrated in Table 3. All classification accuracies are measured on translated images with our pre-trained facial attributes classification model. In terms of classification accuracies, our method surpasses the StarGAN baseline by 3.11%, 5.79% and 2.16% points for hair color, gender and age respectively.

6 Conclusion and Future Work

In this paper, we introduced the concept of deliberation learning for image translation, which shares high-level sense with human behaviors: reviewing and keeping polishing. We implemented deliberation learning on image-to-image translation. The experimental results demonstrate that our method generates more natural images and preserves crucial details.

There are many interesting directions to explore in the future. First, we will apply the idea of deliberation learning to more tasks like supervised image classification and segmentation. Second, deliberation learning can be iteratively used to keep polishing an image to get a better output. Third, we will study how to speed up the algorithm in the future.

Acknowledgments

This work was supported in part by NSFC under Grant 61571413, 61632001. We thank all the anonymous reviewers for their valuable comments on our paper.

[4] https://github.com/yunjey/stargan

References

[Anoosheh *et al.*, 2018] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[Brock *et al.*, 2019] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, and Munyoung Kim. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2018.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016.

[Donahue *et al.*, 2017] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations (ICLR)*, 2017.

[Dong *et al.*, 2014] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*. Springer, 2014.

[Dumoulin *et al.*, 2017] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Olivier Mastropietro, Alex Lamb, Martin Arjovsky, and Aaron Courville. Adversarially learned inference. In *International Conference on Learning Representations (ICLR)*, 2017.

[Ge *et al.*, 2018] Tao Ge, Furu Wei, and Ming Zhou. Fluency boost learning and inference for neural grammatical error correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1055–1065, 2018.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[Guu *et al.*, 2018] Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes. *Transactions of the Association of Computational Linguistics*, 6:437–450, 2018.

[He *et al.*, 2016] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[Kim *et al.*, 2017] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

[Lin *et al.*, 2018] Jianxin Lin, Yingce Xia, Tao Qin, Zhibo Chen, and Tie-Yan Liu. Conditional image-to-image translation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532. IEEE, 2018.

[Liu and Tuzel, 2016] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015.

[Liu *et al.*, 2018] Alexander Liu, Yen-Chen Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[Lucic *et al.*, 2018] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[Mejjati *et al.*, 2018] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Pumarola *et al.*, 2018] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[Reed *et al.*, 2016] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning (ICML)*, pages 1060–1069, 2016.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Wang *et al.*, 2019] Yiren Wang, Yingce Xia, Tianyu He, Fei Tian, Tao Qin, ChengXiang Zhai, and Tie-Yan Liu. Multi-agent dual learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[Xia *et al.*, 2017] Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[Xu *et al.*, 2018] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[Yi *et al.*, 2017] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *International Conference on Computer Vision (ICCV)*, 2017.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, 2017.