Rethinking the Paradigm of Content Constraints in Unpaired Image-to-Image Translation

Xiuding Cai[1 2], Yaoyao Zhu[1 2], Dong Miao[1 2], Linjie Fu[1 2], Yu Yao[1 2*]

[1] Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu, China
[2] University of Chinese Academic Sciences, Beijing, China
{caixiuding20, zhuyaoyao19, miaodong20, fulinjie19}@mails.ucas.ac.cn, casitmed2022@163.com

Abstract

In an unpaired setting, lacking sufficient content constraints for image-to-image translation (I2I) tasks, GAN-based approaches are usually prone to model collapse. Current solutions can be divided into two categories, reconstruction-based and Siamese network-based. The former requires that the transformed or transforming image can be perfectly converted back to the original image, which is sometimes too strict and limits the generative performance. The latter involves feeding the original and generated images into a feature extractor and then matching their outputs. This is not efficient enough, and a universal feature extractor is not easily available. In this paper, we propose EnCo, a simple but efficient way to maintain the content by constraining the representational similarity in the latent space of patch-level features from the same stage of the encoder and decoder of the generator. For the similarity function, we use a simple MSE loss instead of contrastive loss, which is currently widely used in I2I tasks. Benefits from the design, EnCo training is extremely efficient, while the features from the encoder produce a more positive effect on the decoding, leading to more satisfying generations. In addition, we rethink the role played by discriminators in sampling patches and propose a discriminative attention-guided (DAG) patch sampling strategy to replace random sampling. DAG is parameter-free and only requires negligible computational overhead, while significantly improving the performance of the model. Extensive experiments on multiple datasets demonstrate the effectiveness and advantages of EnCo, and we achieve multiple state-of-the-art compared to previous methods.

Introduction

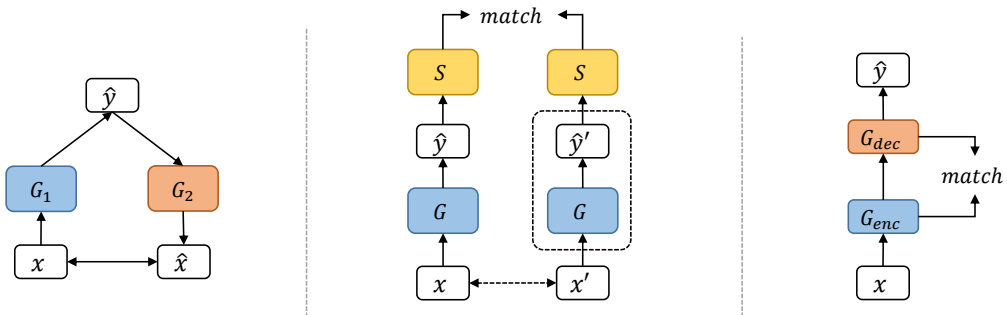
Image-to-image translation (I2I) aims to convert images from one domain to another with content preserved as much as possible. I2I tasks have received a lot of attention given their wide range of applications, such as style transfer (Ulyanov, Vedaldi, and Lempitsky 2016), semantic segmentation (Yu, Koltun, and Funkhouser 2017; Kirillov et al. 2020), super resolution (Yuan et al. 2018), colorization (Zhang, Isola, and Efros 2016), dehazing (Dong et al. 2020) and image restoration (Liang et al. 2021) *etc*.

In an unpaired setting, lacking sufficient content constraints for the I2I task, using an adversarial loss (Goodfellow et al. 2014) alone is often prone to model collapse. To ensure content constraints, current generative adversarial networks (GAN)-based approaches can be broadly classified into two categories. One is reconstruction-based solutions. Typical approaches are CycleGAN (Zhu et al. 2017) and UNIT (Liu, Breuel, and Kautz 2017). They propose the cycle consistency or shared-latent space assumption, which requires that the transformed image, or the transforming image, should be able to map back to the original image perfectly. However, these assumptions are sometimes too strict (Park et al. 2020). For instance, the city street view is converted into a certain pixel-level annotated label, but reconverting a label to a city street view has yet countless possibilities. Such ill-posed setting limits the performance of reconstruction-based GANs, leading to unsatisfactory generations (Chen et al. 2020a).

Another solution for content constraints is Siamese networks (Bromley et al. 1993). Siamese networks are weight sharing neural networks that accept two or more inputs. They are natural tools for comparing entity differences. For the I2I task, the input image and the generated image are fed to some Siamese networks separately, and the content consistency is ensured by matching the output features. CUT (Park et al. 2020) re-exploit the encoder of the generator as a feature extractor and propose the PatchNCE loss to maximize the mutual information between the patches of the input and generated images and achieve superior performance over the reconstruction-based methods. Some studies (Mechrez, Talmi, and Zelnik-Manor 2018; Zheng, Cham, and Cai 2021) repurpose the pre-trained VGG network (Simonyan and Zisserman 2015) as a feature extractor to constrain the feature correlation between the source and generated images. Given the strength and flexibility of Siamese networks, such content-constrained methods are increasingly widely used. However, Siamese network-based GANs mean that the source and generated images need to be fed into the Siamese networks again separately, which entails additional computational costs for training. In addition, an ideal Siamese network that can measure the differences in images well, is not always available.

Can we explicitly constrain the content inside the generator network? Inspired by U-Net (Ronneberger, Fischer,

*Corresponding author.



(a) Reconstruction-based (b) Siamese network-based (c) Ours

Figure 1: A comparison of different content constraints frameworks. (a) Reconstruction-based methods require that $x \leftrightarrow G_2(G_1(x))$, a ℓ_1 loss or ℓ_2 loss is always used. Typical methods are CycleGAN (Zhu et al. 2017), UNIT (Zhu et al. 2017), *etc*. (b) Siamese network-based methods like CUT (Park et al. 2020) or LSeSim (Zheng, Cham, and Cai 2021) complete the content constraint through a defined feature extractor S, *i.e.*, match$(S(G(x), x'))$ or match$(S(G(x), S(G(x'))))$, where x' is the augmented x. Note that the augmentation of x is optional (dashed box). (c) EnCo completes the content constraint by agreeing on the representational similarity of features from the encoder and decoder of the generator.

and Brox 2015), a popular modern network architecture design that integrates features from different stages of the encoder and the decoder by skip connections, we make the encoding-decoding symmetry assumption for the I2I tasks. We assume that the semantic levels of encoder and decoder features from the same stage are the same (note that the number of encoder and decoder stages are opposite).

Based on the assumption, we present EnCo, a simple but efficient way to constrain the content by agreeing on the representational similarity in the latent space of features from the same stage of the **En**coder and de**Co**der of the generator. Specifically, we map the multi-stage intermediate features of the network to the latent space through a projection head, in which we aim to bring closer the representation of the same-stage features from the encoder and decoder, respectively. To prevent the networks from falling into a collapse solution, where the projection learns to output constants to minimize the similarity loss, we stop the gradient of the feature branch from the encoder, as well as add a prediction head for the decoder branch. It is worth mentioning that we find that negative samples are not necessary for the EnCo framework for the content constraint. As a result, we use a simple mean squared error loss instead of the contrastive loss that is widely used in current I2I tasks. This eludes the problems associated with negative sample selection (Hu et al. 2022).

Benefits from the design, the training of EnCo is efficient and lightweight, as the content constraint is accomplished inside the generative network and no reconstruction or Siamese networks are required. To train more efficiently, similar to CUT, we sample patches from the intermediate features of the generative network, and perform patch-level features matching instead of the entire feature map-level. An ensuing question is from which locations we sample the patches. A simple approach is random sampling, which has been adopted by many methods (Park et al. 2020; Han et al. 2021; Zheng, Cham, and Cai 2021). However, this may not be the most efficient (Hu et al. 2022). We note that the discriminator provides key information for the truthfulness of the generated images. However, most of the current GAN-based approaches ignore the potential role of the discriminator in sampling patches (if involved). To this end, we propose a parameter-free discriminative attention-guided patch sampling strategy (DAG). DAG takes advantage of the discriminative information provided by the discriminator and attentively selects more informative patches for optimization. Experimentally, we show that our proposed patch sampling strategy can accelerate the convergence of model training and improve the model generation performance with almost negligible computational effort.

- We propose EnCo, a simple yet efficient way for content constrains by agreeing on the representational similarity of features from the same stage of the encoder and decoder of the generator in the latent space.
- We rethink the potential role of discriminators in patch sampling and propose a parameter-free DAG sampling strategy. DAG improves the generative performance significantly while only requiring an almost negligible computational cost.
- Extensive experiments on several popular I2I benchmarks reveal the effectiveness and advantages of EnCo. We achieve several state-of-the-art compared to previous methods.

Related Works

Image-to-Image Translation Image-to-image translation (Isola et al. 2017; Wang et al. 2018; Zhu et al. 2017; Park et al. 2020; Wang et al. 2021) aims to transform images from the source domain to the target domain with the semantic content preserved. Pix2Pix (Isola et al. 2017) was the first framework to accomplish the I2I task using paired data with an adversarial loss (Goodfellow et al. 2014) and a reconstruction loss. However, paired data across domains are infeasible to be collected in most settings. Methods such as CycleGAN (Zhu et al. 2017), DiscoGAN (Kim et al. 2017)

and DualGAN (Yi et al. 2017) extend the I2I task to an unsupervised setting based on cycle consistency assumption that the generated image should be able to be converted back to the original image again. In addition, UNIT (Liu, Breuel, and Kautz 2017) and MUNIT (Huang et al. 2018) propose to learn a shared-latent space in which the hidden variable, *i.e.*, the encoded image, can be decoded as both the target image and the original image. These methods can be classified as reconstruction-based solutions, and they implicitly assume that the process of conversion should be able to reconvert to the original image. However, perfect reconstruction is unlikely to be possible in many cases, which can potentially limit the performance of the generative networks (Park et al. 2020). In addition, such methods usually require additional auxiliary generators and discriminators.

Siamese Networks for Content Constraints Another solution for content constraints can be attributed to Siamese network-based approaches, and they can effectively address the challenges posed by reconstruction-based ones. Siamese networks usually consist of networks with shared weights that accept two or more inputs, extract features, and compare differences. The selection of Siamese networks for content constraints can be different. For instance, the Siamese network of DistanceGAN and GcGAN is their generator. They require that the distances between the input images and the distances between the output images after generation should be consistent. CUT reuses the encoder of the generator as the Siamese network and proposes PatchNCE loss, aiming to maximize the mutual information between the patches of the input and output images. DCL (Han et al. 2021) extends CUT to a dual-way settings that exploiting two independent encoders and projectors for input and generated images respectively, but doubles the number of network parameters. Some recent studies have also attempted to re-purposed the pre-trained VGGNet (Simonyan and Zisserman 2015) as a perceptual loss to require that the input and output images should be visually consistent (Zheng, Cham, and Cai 2021). There may be a priori limitations in these methods, such as the frozen network weights of the loss function cannot adapt to the data and thus may not be the most appropriate (Zheng, Cham, and Cai 2021). Our work is quite different from the current approaches, as shown in Figure 1, where we impose constraints inside the generative network, *i.e.*, between the encoder and decoder, without requiring additional networks for reconstruction or feature extraction. Therefore, EnCo has a higher training efficiency.

Contrastive Learning Recently, contrastive learning (CL) has achieved impressive results in the field of unsupervised representation learning (Hjelm et al. 2018; Chen et al. 2020b; He et al. 2020; Henaff 2020; Oord, Li, and Vinyals 2018). Based on the idea of discriminative, CL aims to bring closer the representation of two correlated signals (known as positive pair) in the embedding space while pushing away the representation of uncorrelated signals (known as negative pair). CUT first introduced contrastive learning to the I2I task and has been continuously improved since then (Han et al. 2021; Hu et al. 2022; Zhan et al. 2022). QS-Attn (Hu et al. 2022) improved the negative sampling strategy of CUT, by computing the QKV matrix to dynamically selects relevant anchor points as positive and negatives. MoNCE (Zhan et al. 2022) proposed modulated noise contrastive estimation loss to re-weight the pushing force of negatives adaptively according to their similarity to the anchor. However, the performance of CL-based GANs approaches is still affected by the negative sample selection and poor negative may lead to slow convergence and even counter-optimization (Robinson et al. 2020). Therefore, some studies have raised the question whether using of the negative is necessary. BYOL (Grill et al. 2020) successfully trained a discriminative network using only positive pairs with a moment encoder. SimSiam (Chen and He 2020) pointed out that the stopping gradient is an important component for successful training without negatives, thus removing the moment encoder. EnCo is trained without negatives and only considerate the same stage features from the encoder and decoder of the generator as a positive pair, ensuring content consistency.

Methods

Main Idea

Given an image from the source domain $x \in \mathcal{X}$, our goal is to learn a mapping function (also called a generator) $G_{\mathcal{X}\to\mathcal{Y}}$ that converts the image from the source domain to the target domain $\mathcal{Y}$, *i.e.*, $\hat{y} = G_{\mathcal{X}\to\mathcal{Y}}(x)$, and with as much content semantic information preserved as possible.

Traditional content constraint methods based on Siamese networks, such as CUT, intend to constrain the content consistency of an image after generation with the source image. EnCo aims to constrain the content consistency of features generated in the intermediate process from the source image to the target image. Our approach is more efficient to train and achieves better performance. The overall architecture is shown in Figure 2 and contains the generator G, the discriminator D, the projection head h, and the prediction head g. We decompose the generator into two parts, the encoder and the decoder, each of which consists of L-stage sub-networks. For any input source domain image x, after passing through L-stage sub-networks of the encoder, a sequence of features of different semantic levels are produced, *i.e.*, $\{h_l\}_1^L = \{G_{enc}^l(h_{l-1})\}_1^L$. where $x = h_0$. Then feeding the output of the last stage of the encoder, *i.e.*, h_L, into the decoder, we can also obtain a sequence of features of different semantic levels in the decoder $\{h_l\}_{L+1}^{2L} = \{G_{dec}^l(h_{l-1})\}_{L+1}^{2L}$, where $\hat{y} = h_{2L}$. For the I2I task, we make the encoding-decoding symmetry assumption that the semantic levels of the encoder and decoder features f_l and f_{2L-l} from the same stage are the same (note that the number of stages of the encoder and decoder is opposite). For brevity, we abbreviate $2L - l$ as $\tilde{l}$ and denote $(f_l, f_{\tilde{l}})$ as a pair of same-stage features in the following.

We consider that the content of the transformed image can be preserved by constraining $(f_l, f_{\tilde{l}})$. However, this direct approach may degrade the optimization of the generative network. With this view, we propose to guarantee the content constraint by constraining the representational similarity of the encoder and decoder features of the generator in the latent space.

As shown in Figure 2, for any pair of same-stage features

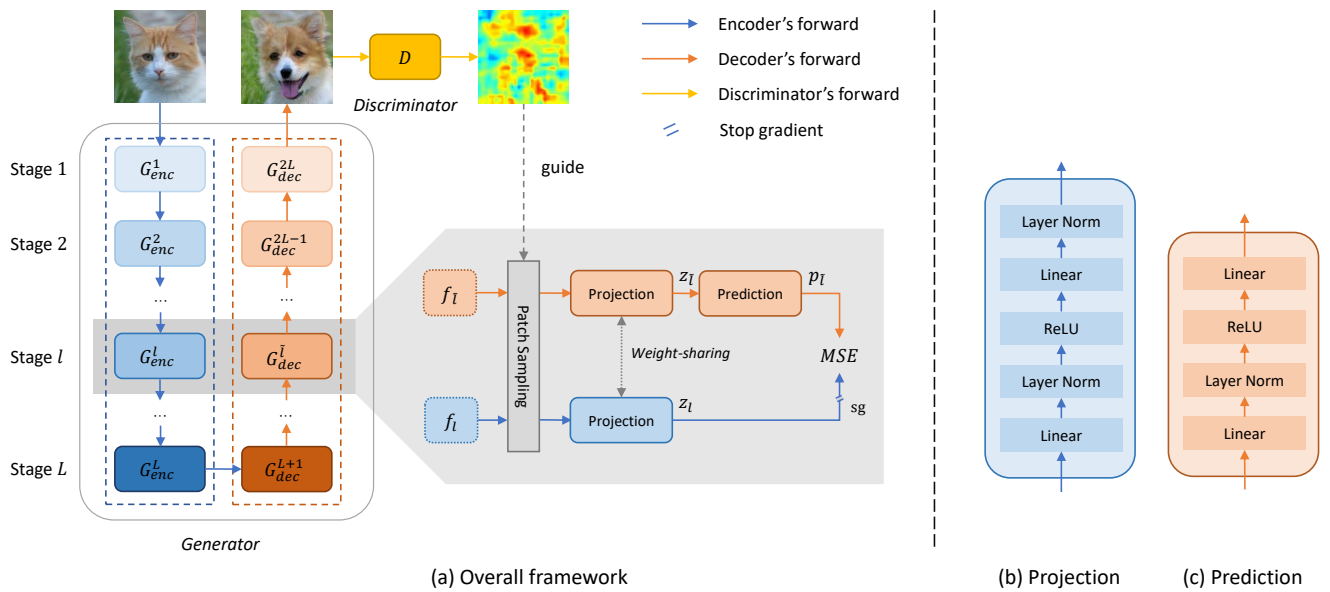


Figure 2: (a) The overview of EnCo framework. EnCo constrain the content by agreeing on the representational similarity in the latent space of features from the same stage of the encoder and decoder of the generator. (b) The architecture of the projection. (c) The architecture of the prediction.

$(f_l, f_{\tilde{l}})$, we map them to the K-dimensional latent space by a shared two-layer projection head $h(\cdot)$ to obtain $z_l \triangleq h(f_l)$ and $z_{\tilde{l}} \triangleq h(f_{\tilde{l}})$. Inspired by (Grill et al. 2020), we further add a prediction head $g(\cdot)$ to $z_{\tilde{l}}$ to enhance the non-linear expression of $z_{\tilde{l}}$, and obtain $p_{\tilde{l}} \triangleq g(z_{\tilde{l}})$.

To avoid collapsing or expanding, we ℓ_2-normalize both z_l and $p_{\tilde{l}}$ and map them to the unit sphere space to obtain $\bar{z}_l \triangleq z_l/\|z_l\|_2$ and $\bar{p}_{\tilde{l}} \triangleq p_{\tilde{l}}/\|p_{\tilde{l}}\|_2$. Finally, we define the following mean-squared error loss aiming to constrain the representational similarity of a pair of same-stage normalized hidden variables from the encoder and decoder,

$$\mathcal{L}(z_{\tilde{l}}, z_l) = \|\bar{p}_{\tilde{l}} - \bar{z}_l\|_2^2 = 2 - 2 \cdot \frac{\langle g(z_{\tilde{l}}), z_l \rangle}{\|g(z_{\tilde{l}})\|_2 \cdot \|z_l\|_2}. \quad (1)$$

To prevent a collapse solution, *i.e.*, the networks learn to output constants to minimize the loss. Following (Chen and He 2020), we solve this problem by introducing the key component of stopping gradient. We modify Eq. (1) as follows,

$$\mathcal{L}(z_{\tilde{l}}, \texttt{stopgrad}(z_l)). \quad (2)$$

This means that during optimization, z_l is constant and $z_{\tilde{l}}$ is expected to be able to predict z_l through the prediction head $g(\cdot)$. Therefore, $z_{\tilde{l}}$ cannot vary too much from z_l, in such a way that the content constraint is achieved.

Multi-stage, Patch-based Content Constraints

Consider a more efficient training approach, given features pair $(f_l, f_{\tilde{l}})$, we sample S patches from different positions of $f_{\tilde{l}}$, feed to the projection and prediction head, and obtain the set $\mathbf{q}_{\tilde{\mathbf{l}}} = \{q_{\tilde{l}}^{(1)}, \cdots, q_{\tilde{l}}^{(S)}\}$, where the subscript indicates which stage to sample and the superscript denotes where to sample from the feature map. Similarly, we can sample from the same position, from f_l, and feed to the projection to get $\mathbf{k}_\mathbf{l} = \{k_{\tilde{l}}^{(1)}, \cdots, k_l^{(S)}\}$. We implement content constraints on patch-level features, rather than the entire feature map-level. Therefore, we just need to use $\mathbf{k}_\mathbf{l}, \mathbf{q}_{\tilde{\mathbf{l}}}$ to replace $z_l, z_{\tilde{l}}$ in Eq. (2), respectively, and get

$$\mathcal{L}(\mathbf{q}_{\tilde{\mathbf{l}}}, \texttt{stopgrad}(\mathbf{k}_\mathbf{l})). \quad (3)$$

We can further extend Eq. (3) to a multi-stage version, *i.e.*,

$$\mathcal{L}_{\text{MultiStage}}(G, h, g, \mathbf{X}) = \mathbb{E}_{x\sim\mathbf{X}} \sum_{l}^{\mathbb{L}} \sum_{s}^{\mathbb{S}_l} \mathcal{L}(q_{\tilde{l}}^{(s)}, \texttt{stopgrad}(k_l^{(s)})), \quad (4)$$

where, $\mathbb{L}$ is the set of chosen same-stage pairwise features to calculate the mean-squared error loss, and $\mathbb{S}_l$ is the set of sampled positions of patches from $(f_l, f_{\tilde{l}})$.

Discriminative Attention-guided Patch Sampling Strategy

We further propose an efficient discriminative attention-guided (DAG) patch sampling strategy to replace the current widely used random sampling strategy used in Eq. (4). The idea of DAG is simple. DAG mainly takes good advantage of the important information from the discriminator: the truthfulness of the generated images, and attentively selects more informative patches for optimization.

Assuming that a total of K patches would be sampled, for any pairwise features $(f_l, f_{\tilde{l}})$, DAG proceeds as follows: 1) obtaining the attention scores: interpolating the output of the discriminator to the same resolution size as f_l and $f_{\tilde{l}}$, thus each position on f_l receives a attention score; 2) over-sampling: uniformly sampling $kK(k > 1)$ patches from $f_{\tilde{l}}$,

where k is the oversampling ratio; 3) ranking: sorting all sampled patches in ascending order according to their corresponding attention scores; 4) importance sampling: selecting the top $\beta K (0 \leq \beta \leq 1)$ patches with the highest scores, where β is the importance sampling ratio; 5) Covering: uniformly sample the remaining $(1-\beta)K$ patches. Note that the DAG is parameter-free, while requiring only an almost negligible computational cost.

Full Objective

In addition to the MultiStage loss presented above, we also use an adversarial loss to complete the domain transfer, and we add an identity mapping loss as a regularization term to stabilize the network training.

Generative adversarial loss We use LSGAN loss (Mao et al. 2017) as the adversarial loss to encourage the generated images that are as visually similar to images in the target domain as possible, which is formalized as follows,

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{y\sim Y}[D(y)^2] + \mathbb{E}_{x\sim X}[(1 - D(G(x)))^2]. \quad (5)$$

Identity mapping loss In order to stable the training and accelerate the convergence, we add an identity mapping loss.

$$\mathcal{L}_{\text{identity}}(G) = \mathbb{E}_{y\sim Y}\|G(y) - y\|_1. \quad (6)$$

We only use this regular term in the first half of the training phase because we find that it impacts the generative performance of the network to some extent.

Overall loss Our final objective function is as follows:

$$\begin{aligned}\mathcal{L}_{\text{total}}(G, D, h, g) = &\mathcal{L}_{\text{GAN}}(G, D, X, Y) \\ &+ \lambda_{NCE}\mathcal{L}_{\text{MultiStage}}(G, h, g, X) \\ &+ \lambda_{IDT}\mathcal{L}_{\text{identity}}(G, Y),\end{aligned} \quad (7)$$

where λ_{NCE} and λ_{IDT} are set to 2 and 10, respectively.

Discussion

EnCo achieves content consistency by constraining the similarity between representations of the encoder and decoder features at multiple stages in the latent space. In fact, there are two additional perspectives on how EnCo achieves content consistency that we would like to offer. Firstly, in relation to reconstruction-based methods, EnCo requires decoding features that should be able to predict their corresponding encoded features in turn through the prediction MLP, which is somehow similar to the reconstruction-based approach. *However, EnCo conducts the reconstruction at the feature level rather than the pixel level, which provides more freedom when enforcing content consistency.* EnCo degenerates into a special CycleGAN approach when we only constrain the consistency of the input and generated images, with projection being an identity network and prediction network being another generator. Secondly, EnCo can also be regarded as an implicit and lightweight Siamese network paradigm (similar to CUT), where encoder and decoder features are constrained to have similar representations in the latent space through the shared projection MLP.

Experiments

Experiment Setup

Datasets To demonstrate the superiority of our method, we trained and evaluated on three popular I2I benchmark datasets, including *Cityscapes*, *Cat→Dog*, *Horse→Zebra*. *Cityscapes* (Cordts et al. 2016) contains street scenes from German cities with 2975 training images and 500 test images. Each image in *Cityscapes* has a resolution of 2048×1024 with high quality pixel-level annotation. *Cat→Dog* comes from the AFHQ dataset (Choi et al. 2020), containing 5153 images of cats and 4739 images of dog. *Horse→Zebra*, collected by CycleGAN from ImageNet (Deng et al. 2009), contains 1067 images of horses and 1334 images of zebras. For all experiments, we resized images to 256×256 resolution size.

Implementation details We use the same ResNet-based generator (Park et al. 2020) and PatchGAN discriminator (Isola et al. 2017) with a receptive field of 70×70 as CUT. We use Adam optimizer (Kingma and Ba 2014) with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For the *CityScapes* and *Horse→Zebra* datasets, 400 epoches are trained, and 200 epoches are trained only for the *Cat→Dog* dataset. Following TTUR (Heusel et al. 2017), we set unbalanced learning rates of $5e-5$, $2e-4$ and $5e-5$ for the generator, discriminator and projection head, respectively. We start linearly decaying the learning rate halfway through the training with batch size of 1. We use by default a two-layer projection and a two-layer prediction with a dimension of 256 for all linear layers, the same as used in the CUT, but we add layer normalization after the linear layers, except for the last linear layer of the prediction (See Figure 2 (b) and (c)).

Evaluation metrics We mainly use Fréchet Inception Distance (FID) (Heusel et al. 2017) to evaluate the visual quality of the generated images. FID is one of the most commonly used distribution-based image quality assessment metrics, by comparing the distance between distributions of generated and real images in a deep feature domain. For the *CityScapes* dataset, since they have corresponding segmentation labels, following CUT, we apply a pre-trained segmentation model (Yu, Koltun, and Funkhouser 2017) to segment the generated images and use three evaluation metrics, including mean average precision (mAP), pixel-wise accuracy (pixAcc), and average class accuracy (classAcc) to measure how well methods discover correspondences. In addition, we also include the training speed and memory consumption to measure the training efficiency of the model.

Comparison with the State-of-the-art Methods

We compare our method with several state-of-the-art methods of unpaired I2I, including CycleGAN (Zhu et al. 2017), MUNIT (Huang et al. 2018), GcGAN (Fu et al. 2019), CUT (Park et al. 2020), DCLGAN (Han et al. 2021), FSeSim (Zheng, Cham, and Cai 2021), MoNCE (Zhan et al. 2022) and QS-Attn (Hu et al. 2022). Among them, CycleGAN and MUNIT are reconstruction-based methods, and the rest are Siamese network-based ones.

Quantitative and qualitative results The results of the quantitative comparisons are shown in Table 1. As can be

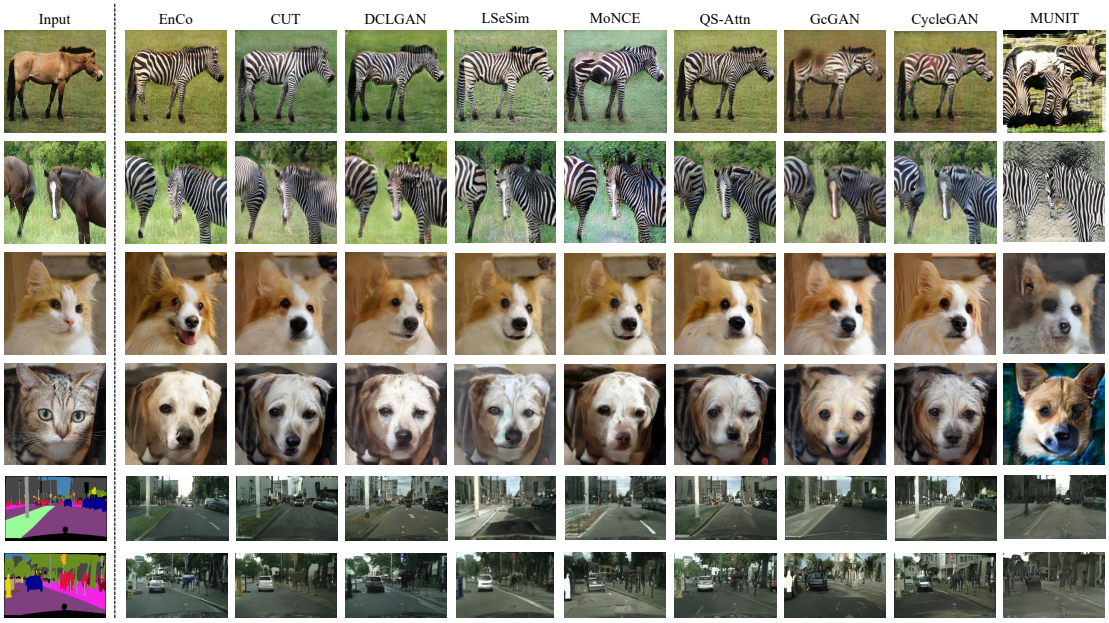


Figure 3: Results of qualitative comparison. We compare EnCo with existing methods on the *Horse→Zebra*, *Cat→Dog*, and *Cityscapes* datasets. EnCo achieves more satisfactory visual results. For example, in the case of *Cat→Dog*, EnCo generates a clearer nose for the dog. And in the case of *Cityscapes*, EnCo successfully generates the traffic cone represented in yellow in the semantic annotation, while the other methods yielded only suboptimal results.

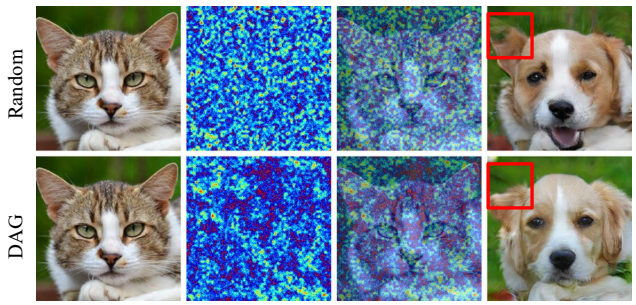
Method	**CityScapes**				**Cat→Dog**	**Horse→Zebra**		
	mAP↑	**pixAcc↑**	**classAcc↑**	**FID↓**	**FID↓**	**FID↓**	**Mem(GB)↓**	**sec/iter↓**
CycleGAN (Zhu et al. 2017)	20.4	55.9	25.4	68.6	85.9	66.8	4.81	0.40
MUNIT (Huang et al. 2018)	16.9	56.5	22.5	91.4	104.4	133.8	3.84	0.39
GCGAN (Fu et al. 2019)	21.2	63.2	26.6	105.2	96.6	86.7	**2.67**	0.62
CUT (Park et al. 2020)	24.7	68.8	30.7	56.4	76.2	45.5	3.33	0.24
DCLGAN (Han et al. 2021)	22.9	77.0	29.6	49.4	60.7	43.2	7.45	0.41
FSeSim (Zheng, Cham, and Cai 2021)	22.1	69.4	27.8	54.3	87.8	43.4	2.92	0.17
MoNCE (Zhan et al. 2022)	25.6	78.4	33.0	54.7	74.2	41.9	4.03	0.28
QS-Attn (Hu et al. 2022)	25.5	**79.9**	31.2	53.5	72.8	41.1	2.98	0.35
EnCo	**28.4**	77.3	**37.2**	**45.4**	**54.7**	**38.7**	2.83	**0.14**

Table 1: Comparison with baselines on unpaired image translation. We compare our approach to the state-of-the-art methods on three datasets. We show multiple metrics, where the ↑ indicates higher is better and the ↓ indicates lower is better. It is worth noting that our method outperforms all baselines on the FID metric and shows superior results on the *Cityscapes* for the semantic segmentation metric. Also, our method shows a fast training speed.

seen from Table 1, EnCo outperforms all baselines on FID metric on three datasets. In particular, on tasks *Cityscapes* and *Cat→Dog*, EnCo's FID drops by 4 and 6 points, respectively, compared to the second place (indicated by underlining). Figure 3 gives a qualitative comparison. It can be seen that EnCo achieves more satisfactory generation results. For example, in the case of *Cat→Dog*, EnCo succeeds in generating a clear dog's nose, but most other methods generate only suboptimal results.

EnCo also achieved the highest scores in terms of segmentation indicators on the *Cityscapes*, except for the pixAcc indicator, which is slightly below QS-Attn. This may come from the problem of class imbalance. However, it is worth noting that EnCo is much higher than the other methods in the classAcc metric (37.2 compared to the second place 33.0). This indicates that EnCo uncovers more category-pixel correspondences in the generation. For example, in the last row of Figure 3, EnCo successfully generates the traffic cone represented in yellow in the semantic annotation, while all the other methods fail.

Model efficiency analysis We also report the training speed and memory consumption of the different models in Table 1. As can be seen, EnCo exhibits fast training speed with sec/iter 0.14, which is 2.9× faster than the reconstruction-



(a) Input x (b) Frequency (c) Frequency on x (d) Generated $\hat{y}$

Figure 4: Comparison of different patch sampling strategy. For the input image x, we superimpose the sampling positions every 10 epochs to obtain the sampling frequency map (b). As can be seen from (c), compared to the random strategy, our proposed DAG patch sampling strategy are more focused on regions that help in domain discrimination, such as ears, eyes, nose, *etc*. As a result, the model with DAG sampling strategy generates more adorable results than random sampling strategy (see the red box in (d)).

	Configurations	Cat→Dog	Horse→Zebra	
		FID ↓	FID ↓	sec/iter
A	(h_7, h_{24})	63.8	43.1	**0.129**
B	$(h_7, h_{24}), (h_{13}, h_{18})$	63.8	43.1	0.134
C	EnCo†	55.9	39.2	0.136
D	uniform ($k = 1, \beta = 0.0$)	63.8	43.1	0.135
E	default ($k = 4, \beta = 0.5$)	**54.7**	**38.7**	0.136

Table 2: Ablations on oversampling ratio k and importance sampling ratio β of DAG patch sampling strategy.

based approach of CycleGAN and 1.7× faster than the Siamese network-based approach of CUT. Meanwhile, the memory usage of EnCo is extremely efficient, with only 2.83 GB compared to the lowest one 2.67 GB. The training efficiency of EnCo comes from its content-constrained design, which requires neither networks for reconstruction nor feature extraction, such as Siamese networks.

Ablation Study

Compared to baselines, our method exhibits superior performance. We design several ablation experiments to analyze each contribution of components in isolation. We mainly conduct ablation experiments on two datasets, *Cat→Dog* and *Horse→Zebra*. We report the results of ablation experiments in Table 2.

Influence of multi-scale pairwise features constraint. We initially investigated the impact of multi-scale pairwise features on generative performance. As shown in rows A-B of Table 2, it is evident that the model's performance improves to varying degrees with the inclusion of additional pairs of same-scale features with different semantic levels.

Impact of asymmetric generator architecture. We assume the process of encoding and decoding is symmetric rather than the network design, which means that EnCo can be

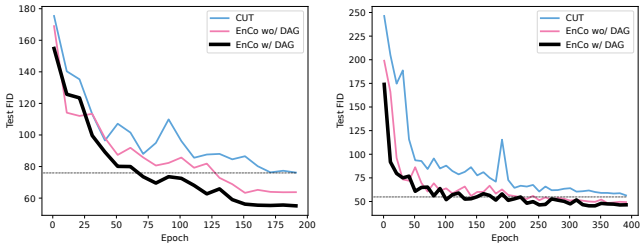


Figure 5: Comparison of test FIDs over training time in tasks *Cat→Dog* (left) and *Cityscapes* (right).

used for asymmetric generators. To prove this, we added a new experiment (EnCo†) in Table 2, which simulates asymmetry. For compatibility, we use bilinear interpolation and add a simple pre-projection MLP to align a feature pair. Specifically, we selected three non-symmetric feature pairs, (h_3, h_{24}), (h_7, h_{20}), and (h_{13}, h_{17}), with h_3 having a resolution of 256 and 64 channels, while h_{24} has a resolution and number of channels of 128. As shown in Table 2, due to the deliberately asymmetric design, EnCo† exhibits a slight performance degradation compared to EnCo, but it still maintains its robustness and outperforms most of the baselines.

Effectiveness of the DAG strategy. We also conducted the ablation experiments for the DAG sampling strategy. As can be seen from row D in Table 2, when we remove the DAG sampling strategy, *i.e.*, use the random strategy, the performance of the model on each task shows different degrees of degradation compared to the default setting, and training speed was hardly affected, which fully demonstrates the effectiveness of DAG patch sampling strategy. Figure 4 gives an example of how DAG negative sampling strategy affects the generating results. We further visualize in Figure 5 showing how the test FID changes with training. The results show that without/with DAG, EnCo reached full CUT performance at the 130/65-th (left) and 135/80-th (right) epochs, respectively. This demonstrates well that the DAG strategy accelerates model convergence.

Conclusion

In this paper, we introduce EnCo, a novel framework for image-to-image tasks. EnCo preseveres content by agreeing on the representational similarity in the latent space of features from the same stage of the encoder and decoder of generator. Compared to current reconstruction-based or Siamese network-based methods, EnCo offers more efficient training, as the constraints are integrated within the generative model. We also rethink the role of the discriminator in the patch sampling and propose a parameter-free discriminative attention-based patch sampling strategy, which incurs almost negligible computational overhead while significantly enhancing the generative performance . Through extensive experiments on multiple popular datasets, we have demonstrated the efficiency and effectiveness of EnCo. We hope EnCo will bring some new thoughts and inspiration to the paradigm of content constrains for unpaired I2I tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 82073338, and in part by the Sichuan Provincial Science and Technology Department under Grant 2022YFS0384 and 2022YFQ0108.

References

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Chen, R.; Huang, W.; Huang, B.; Sun, F.; and Fang, B. 2020a. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8168–8177.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*.

Chen, X.; and He, K. 2020. Exploring Simple Siamese Representation Learning. (arXiv:2011.10566).

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 248–255. Ieee.

Dong, Y.; Liu, Y.; Zhang, H.; Chen, S.; and Qiao, Y. 2020. FD-GAN: Generative adversarial networks with fusion-discriminator for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10729–10736.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Zhang, K.; and Tao, D. 2019. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2427–2436.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P. H.; Buchatskaya, E.; Doersch, C.; Pires, B. A.; Guo, Z. D.; Azar, M. G.; Piot, B.; Kavukcuoglu, K.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. (arXiv:2006.07733).

Han, J.; Shoeiby, M.; Petersson, L.; and Armin, M. A. 2021. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 746–755.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 9729–9738.

Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *ICML*, 4182–4192. PMLR.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*.

Hu, X.; Zhou, X.; Huang, Q.; Shi, Z.; Sun, L.; and Li, Q. 2022. QS-Attn: Query-Selected Attention for Contrastive Learning in I2I Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18291–18300.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning (ICML)*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations (ICLR)*.

Kirillov, A.; Wu, Y.; He, K.; and Girshick, R. 2020. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9799–9808.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1833–1844.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems (NIPS)*, 700–708.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mechrez, R.; Talmi, I.; and Zelnik-Manor, L. 2018. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, 768–783.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision (ECCV)*.

Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Simonyan, K.; and Zisserman, A. 2015. Very deep convolutional networks for large-scale image recognition.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2016. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, W.; Zhou, W.; Bao, J.; Chen, D.; and Li, H. 2021. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14020–14029.

Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *IEEE International Conference on Computer Vision (ICCV)*, 2849–2857.

Yu, F.; Koltun, V.; and Funkhouser, T. 2017. Dilated residual networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, 472–480.

Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; and Lin, L. 2018. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 701–710.

Zhan, F.; Zhang, J.; Yu, Y.; Wu, R.; and Lu, S. 2022. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18280–18290.

Zhang, R.; Isola, P.; and Efros, A. A. 2016. Colorful image colorization. In *European conference on computer vision (ECCV)*, 649–666. Springer.

Zheng, C.; Cham, T.-J.; and Cai, J. 2021. The spatially-correlative loss for various image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16407–16417.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision (ICCV)*.