Vector Quantized Image-to-Image Translation

Yu-Jie Chen[⋆,1,2], Shin-I Cheng[⋆,1,2], Wei-Chen Chiu[1,2],
Hung-Yu Tseng[3], and Hsin-Ying Lee[4]

[1]National Chiao Tung University, Taiwan [2]MediaTek-NCTU Research Center [3]Meta
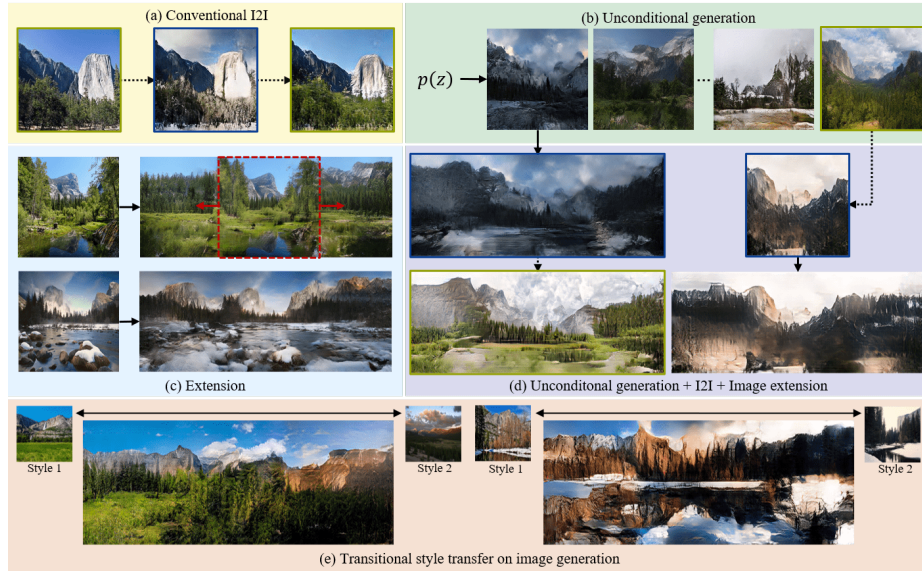[4]Snap Inc.

Fig. 1. Applications of Vector Quantized Image-to-Image Translation. Our proposed method enables several applications: (a) conventional image-to-image translation, (b) unconditional image generation, (c) image extension, (d) arbitrary combination of aforementioned operations, *e.g.* translation and extension on unconditionally generated images, and (e) image generation with transitional stylization. Here we use green frame for summer images and blue frame for winter images.

Abstract. Current image-to-image translation methods formulate the task with conditional generation models, leading to learning only the recolorization or regional changes as being constrained by the rich structural information provided by the conditional contexts. In this work, we propose introducing the vector quantization technique into the image-to-image translation framework. The vector quantized content representation can facilitate not only the translation, but also the unconditional

---

⋆ Equal contribution.

Project page: https://cyj407.github.io/VQ-I2I/

distribution shared among different domains. Meanwhile, along with the disentangled style representation, the proposed method further enables the capability of image extension with flexibility in both intra- and inter-domains. Qualitative and quantitative experiments demonstrate that our framework achieves comparable performance to the state-of-the-art image-to-image translation and image extension methods. Compared to methods for individual tasks, the proposed method, as a unified framework, unleashes applications combining image-to-image translation, unconditional generation, and image extension altogether. For example, it provides style variability for image generation and extension, and equips image-to-image translation with further extension capabilities.

Keywords: Image-to-Image Translation, Vector Quantization, Image Synthesis, Generative Models

1   Introduction

Image-to-image translation (I2I) aims to learn the mapping between different visual domains. Upon being formulated as a conditional generation problem, I2I methods can tackle translation with either paired [13] or unpaired data [34], and perform diverse translations by disentangling the content and style factors of each input domain [12,17,36]. These I2I methods unleash various applications, such as style transfer [11], synthesis from semantic map or layout [2,10,29,32], domain adaptation [6,24], and super-resolution [16].

Most existing I2I methods model the task as a pixel-level conditional generation problem. However, as the conditional contexts are already informative in structure and details, the translation tends to learn simple recolorization or regional transformation without understanding the real target distribution. Is it possible to jointly learn the translation as well as the unconditional distribution to fully exploit the data and make both trainings mutually beneficiary? One intuitive formulation is to define a domain-invariant joint latent distribution, then perform domain-specific maximum likelihood estimation on it. Pixel space is a natural option for the joint latent distribution, yet it struggles to scale due to its computational expensive auto-regressive process. Recently, vector quantization (VQ) technique has shown its effectiveness as an intermediate representation of generative models [7,30]. We thus explore in this work the applicability of adopting vector quantization as the latent representation in the I2I task.

We introduce VQ-I2I, a framework that adopts a vector quantized codebook as an intermediate representation which is able to enable both the image-to-image translation and the unconditional generation of input domains. VQ-I2I consists of a joint domain-invariant content encoder, domain-specific style encoders, and domain-specific decoders. The joint content encoder enforces a shared latent distribution among different domains. The encoded content representation can be further decoded with the style representation obtained from the same input for realizing the self-reconstruction or with that from different inputs for achieving intra- and inter-domain translations. Moreover, with different style representations being given, VQ-I2I is also able to perform diverse translations.

In addition to conventional image-to-image translation, we learn an auto-regressive model on the joint quantized content space to unconditionally synthesize the latent content representation. The capability of unconditional content generation with disentangled style representation can unleash several applications: As shown in Figure 1, VQ-I2I has the multifunctionality for performing I2I, unconditional image synthesis, and image extension. Combining these operations, VQ-I2I can achieve extension on generated samples with the flexibility of stylizing into different domains, and image generation with transitional stylization. These cannot be done by a unified framework to the best of our knowledge.

We conduct extensive quantitative and qualitative evaluations. We measure the realism with the Fréchet inception distance (FID) [9] and subjective study using the AFHQ [4], Yosemite [34], and portrait [17] datasets. On the Cityscapes dataset [5], we use FID metric as well to compare with the I2I methods trained upon paired data. Qualitatively, we demonstrate realistic and diverse I2I translation as well as applications including unconditional generation, image extension, completion, transitional stylization, or combinations over them.

2   Related Work

2.1   Image-to-Image Translation.

Image-to-image translation, first addressed in [13] , aims at learning the mapping function between the source and the target domain. Following works focus on tackling two major challenges: how to handle unpaired data and how to model diverse translations. Cycle-consistency is adopted to handle unpaired data [34,20], while augmented attribute space is proposed to provide diversity [35]. Following efforts are made to handle both challenges jointly [12,17,18], one-sided translation without cycle-consistency [28], to improve the diversity [21,22], and to better handle geometric transformations [14]. Take a step forward, we propose a framework that can perform not only cross-domain translations, but also enable unconditional generation and image extension using the learned representation.

2.2   Vector Quantized Generative Models.

Generative models can be roughly divided into two streams: implicit and explicit density estimation methods. Generative adversarial network, the representative of the implicit method, has been dominant due to its high-fidelity synthesized images, yet suffering from instability in training. On the other hand, explicit methods are more tractable in training but limited to relatively blurred outputs (*e.g.* variational autoencoder (VAE) [15]) or in scaling due to the pixel-level auto-regressive process (*e.g.* PixelRNN [25] and PixelCNN [26]). Recently, vector quantization (VQ) technique has adopted explicit methods to alleviate the scaling issue with quantized latent vectors serving as latent representation [27,30,8,33]. VQGAN then proposes a hybrid framework to first leverage GAN technique to learn VQ codebook, then adopt transformer [7] to train an auto-regressive model on the learned VQ indices. In this work, we propose adopting VQ technique in the I2I task.
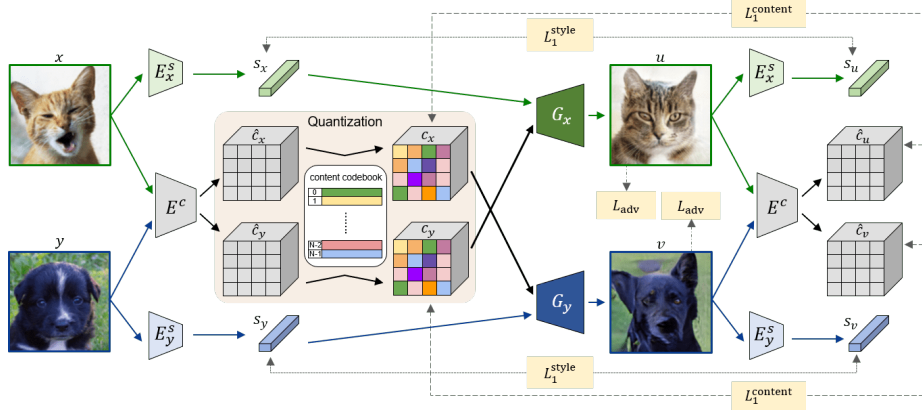
(a) Overall architecture of vector-quantized I2I with disentangled representations.
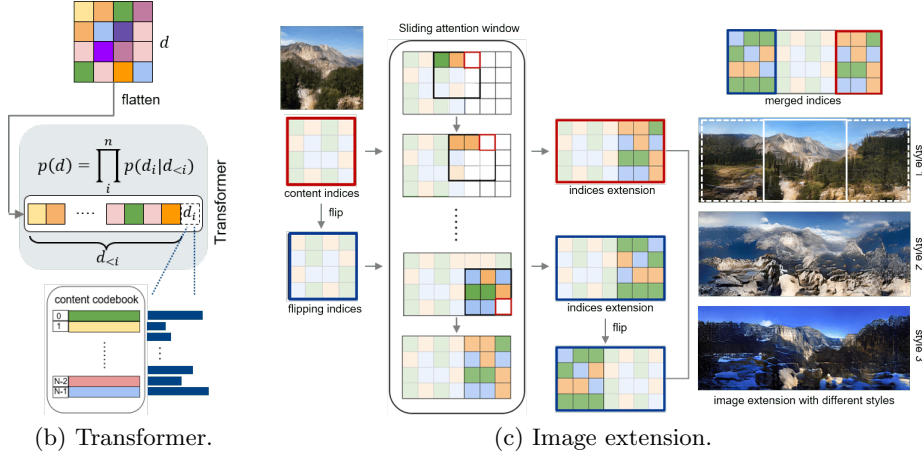

(b) Transformer.

(c) Image extension.

Fig. 2. Method Overview. (a) The proposed framework learns to perform translation with disentangled vector-quantized domain-invariant content and domain-specific style representations. (b) Given the quantized content indices d, we can learn the content distribution in an autoregressive manner using a transformer model. (c) With learned transformer model and the translation model, we can expand an image on both horizontal sides by spatially extending the content map and its flipped one with a sliding attention window. The extended content can be further translated into different styles.

3   Method

As previously motivated, our goal is to leverage the vector quantized codebook, an intermediate representation for 1) image-to-image translation between two visual domains $X \subset \mathbb{R}^{H \times W \times 3}$ and $Y \subset \mathbb{R}^{H \times W \times 3}$ and 2) unconditional generation in each domain. As illustrated in Figure 2 (a), our framework consists of a shared content encoder E^c, a vector quantized content codebook Z, style encoders $\{E_X^s, E_Y^s\}$, generators $\{G_X, G_Y\}$, and discriminators $\{D_X, D_Y\}$. Given an input

image, the content encoder E^c extracts the *vector-quantized* domain-invariant representations, while the style encoders E_X^s, E_Y^s compute the domain-specific features for domain X and Y respectively. The generators G_X, G_Y combine the content representation and style feature to produce the image in each domain. Finally, the discriminators D_X, D_Y aim to distinguish between the generated and real images.

3.1   Vector Quantized Content Representation

Our approach leverages the vector quantization strategy to encode the domain-invariant image content information. Specifically, we construct a codebook $Z = \{z_k\}_{k=1}^K$ that consists of learned content codes $z_k \in \mathbb{R}^{n_c}$, where n_c indicates the code dimension. Given a continuous map $\hat{c} \in \mathbb{R}^{h \times w \times n_c}$ extracted by the content encoder E^c, we find for each spatial entry $\hat{c}_{ij} \in \mathbb{R}^{n_c}$ of $\hat{c}$ its closest code in the codebook Z for obtaining the vector quantized content representation c:

$$c = \mathbf{vq}(\hat{c}) := (\arg\min_{z_k \in Z} \|\hat{c}_{ij} - z_k\|) \in \mathbb{R}^{h \times w \times n_c}. \tag{1}$$

Since the quantization operation $\mathbf{vq}$ is not differentiable for gradient back-propagation, we use the straight-through trick [27] that copies the gradient from c to $\hat{c}$. We learn the codebook Z using the self-reconstruction path and the loss function L_{vq} and L_1^{recon}, where

$$L_{\text{vq}} = \|\text{sg}[\hat{c}] - c\|_2^2 + \|\text{sg}[c] - \hat{c}\|_2^2, \tag{2}$$

where $\text{sg}[\cdot]$ is the stop-gradient operation. We provide the details of the self-reconstruction path and the loss L_1^{recon} later in Section 3.2.

3.2   Diverse Image-to-Image Translation

To enable multi-modal image-to-image translation, our approach learns the disentangled domain-*invariant* content representations and domain-*specific* style features [17,20]. As shown in Figure 2(a), we use an *shared* encoder E^c to extract the content representation for images of two domains, followed by applying the vector quantization operation $\mathbf{vq}$, and use separate encoders E_X^s, E_Y^s to compute the style features:

$$\begin{aligned} c_x, s_x &= \mathbf{vq}(E^c(x)), E_X^s(x) \\ c_y, s_y &= \mathbf{vq}(E^c(y)), E_Y^s(y). \end{aligned} \tag{3}$$

Since the content space is shared among two domains, we can perform the image-to-image translation by swapping the content representations c_x and c_y. Finally, the generators G_X, G_Y use AdaIN normalization layers [11,12] to combine the swapped content representations and style features to synthesize the translated images $u \in X$ and $v \in Y$:

$$u = G_X(c_y, s_x),\ v = G_Y(c_x, s_y). \tag{4}$$

Image-to-Image Translation Training. We use the discriminators D_X and D_Y to impose the domain adversarial loss L_{adv}. The loss L_{adv} encourages the realism of the translated images u for domain X and v for Y.

Nevertheless, training our model with the domain adversarial loss along cannot guarantee the disentanglement of content and style representations. The content map c may encode the style information, thus the generator ignores the style feature s for synthesizing the translated images. To address this issue, we use the latent style regression loss to enforce the bijection between the style features and the translated images:

$$
\begin{aligned}
L_1^{\mathrm{style}} = &\ \|E_X^s(G_X(c_y, s_x)) - s_x\| \\
&+ \|E_Y^s(G_Y(c_x, s_y)) - s_y\|.
\end{aligned}
\tag{5}
$$

We also use the latent content regression loss to facilitate the training:

$$
\begin{aligned}
L_1^{\mathrm{content}} = &\ \|E^c(G_X(c_y, s_x)) - c_y\| \\
&+ \|E^c(G_Y(c_x, s_y)) - c_x\|,
\end{aligned}
\tag{6}
$$

where Figure 2 (a) shows the computation flows behind L_1^{style} and L_1^{content}.

Self-Reconstruction Training. In addition to image-to-image translation, we also involve a self-reconstruction path (*i.e.* reconstructing an image by using its own content and style representations) during the training stage for two empirical reasons. First, self-reconstruction training is vital for learning a meaningful vector-quantized codebook [27]. Second, it facilitates the overall image-to-image training process. Specifically, we impose the self-reconstruction loss:

$$
L_1^{\mathrm{recon}} = \|G_X(c_x, s_x) - x\| + \|G_Y(c_y, s_y) - y\|.
\tag{7}
$$

As described in Section 3.1, we only apply the vector quantization loss L_{vq} (cf. Equation 2) in the self-reconstruction path. The full objective function of our model (L_D for training discriminators; $L_{E^c, Z, E^s, G}$ for training encoders, codebook, and generators) is then summarized as:

$$
\begin{aligned}
L_D &= L_{\mathrm{adv}}, \\
L_{E^c, Z, E^s, G} &= -\lambda_{\mathrm{adv}} L_{\mathrm{adv}} + \lambda_1^{\mathrm{recon}} L_1^{\mathrm{recon}} + \lambda_{\mathrm{vq}} L_{\mathrm{vq}} \\
&\quad + \lambda_1^{\mathrm{content}} L_1^{\mathrm{content}} + \lambda_1^{\mathrm{style}} L_1^{\mathrm{style}},
\end{aligned}
\tag{8}
$$

where λ controls the importance of each loss term. Note that we only optimize the codebook with the vector quantization loss L_{vq} and reconstruction loss L_1^{recon}.

3.3   Unconditional Generation

Vector quantization on the shared content space enables unconditional generation, since we can model the domain-invariant joint (content) distribution using an autoregressive approach [1]. We present the approach in Figure 2 (b). The
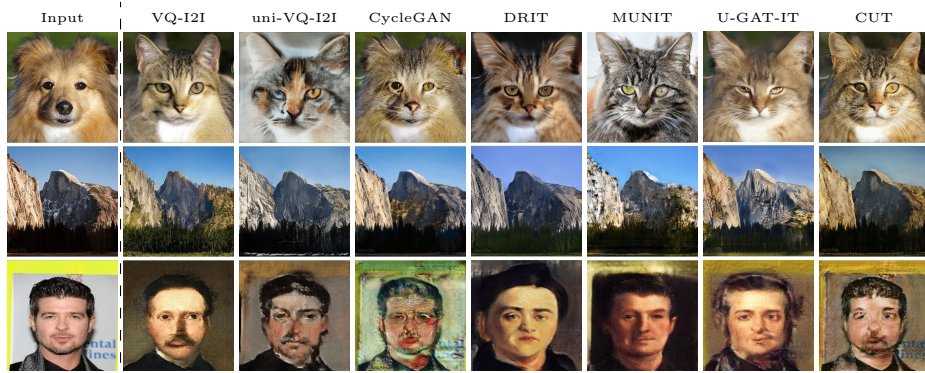
(a) Unpaired I2I Comparison
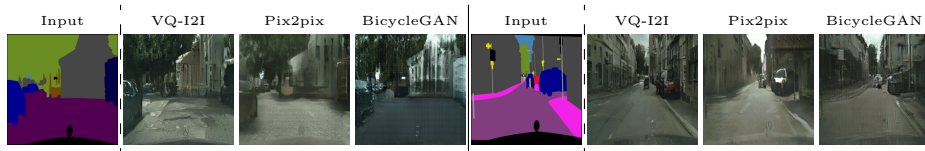
(b) Paired I2I Comparison

Fig. 3. Qualitative Comparisons with Conventional Image-to-Image Translation Methods. (a) We show the translated results of different methods on three unpaired datasets. From top to bottom rows are dog→cat [4], winter→summer [34], and photo→portrait [17]. (b) Our model is able to handle training with paired data on Cityscapes dataset [5]. For each example set (composed of four columns), the leftmost column shows the semantic segmentation of street scenes, and the other columns show the corresponding generated scenes by various models which are trained on paired data.

spatial entries in the content representation c can be represented as a set of indices d in the codebook $Z = \{z_k\}_{k=1}^{K}$, where $c_{ij} = z_{d_{ij}}$. By ordering the index set d using a particular rule, the content generation task can be formulated as the next-index prediction problem. Specifically, given content indices $d_{<i}$, the goal is to predict the distribution of next index d_i: $p(d) = \prod_i p(d_i|d_{<i})$. We train a transformer network [7] for this task by maximizing the log-likelihood of the content representation:

$$L_{\text{transformer}} = \mathbb{E}_{x \sim p(x)}[-\log p(d)]. \tag{9}$$

We provide the ordering details (*i.e.* slight difference between training and testing stages as similar to [7]) in the supplementary materials.

During inference, we first generate the complete content representation using the autoregressive next-index prediction process. Then we combine some style features $\{s_x, s_y\}$, and use the generators $\{G_X, G_Y\}$ to synthesize the image for different domains.
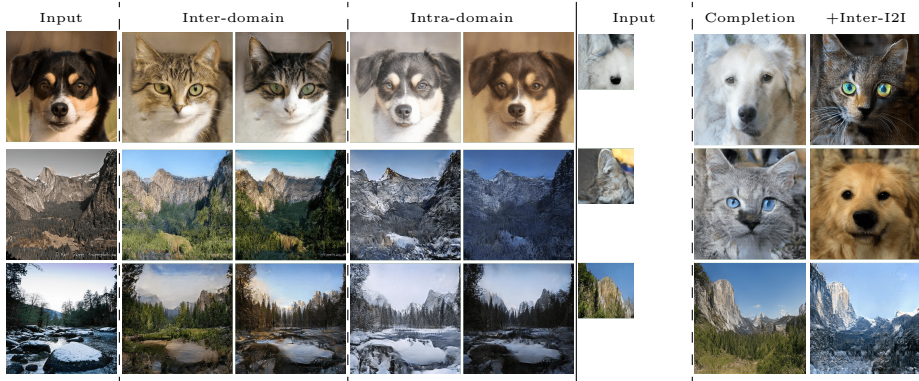
Input	Inter-domain	Intra-domain	Input	Completion	+Inter-I2I

Fig. 4. Diverse Image Translation and Completion. (*left*) We demonstrate both inter-domain and intra-domain translations with the query images (leftmost column) combined with various styles on the dog→cat and winter→summer scenarios. *(right)* Given a quarter of an image from AFHQ [4] or Yosemite [34] dataset as the input, we perform image completion AND the inter-domain translation. VQ-I2I is able to not only learn the joint content distribution of both domains, thus achieving reasonable completion, but support the diverse translation via the design of the disentanglement.

3.4   Content Extension.

Our autoregressive next-index prediction process enables not only the unconditional content generation, but also *content extension*: extending the content of existing images. We illustrate the process in Figure 2 (c). Specifically, given a vector quantized content representation extracted from an existing image, we use the learned transformer model to spatially extend the content map (red outline). By flipping the content representation horizontally, we can extend the content to the opposite direction using the same process (blue outline). The resultant content map which has been extended (on both horizontal sides) then can be gone through generators together with a style feature to produce the extension.

4   Experiments

We evaluate the proposed framework on image translation, unconditional generation and image extension. We compare VQ-I2I with several representative I2I, image generation and outpainting approaches. We then demonstrate various applications of our framework which seamlessly combine I2I with unconditional image generation, image extension, and transitional stylization. Finally, we conduct the ablation study to understand the efficacy of different design choices.

Datasets. We conduct experiments using both paired and unpaired I2I datasets. For unpaired datasets, we use the Yosemite dataset [34] for the shape-invariant translation, and the AFHQ [4] and portrait [17] datasets for the shape-variant translation task. For paired dataset, we use the Cityscapes dataset [5].

Fig. 5. Qualitative Examples on Image Extension. Example results of image extension on Yosemite [34] datasets, where the comparison with respect to Boundless [31] baseline is also provided. The leftmost column shows the input images for the image extension, where the model takes left portion of size 256×256 for each input image and extends for 192 pixel width toward the right-hand side. VQ-I2I is able to generate smooth and diverse extensions with style variability.

Compared Baselines. For the unpaired I2I setting, we compare our method with CycleGAN [34], DRIT [17], MUNIT [12], and recent CUT [28] and U-GAT-IT [14]. For the paired I2I setting, we make a comparison between our method and Pix2pix [13] as well as BicycleGAN [36]. For unconditional generation, we compare our approach with VQGAN [7]. As for image extension [3,19,31], we consider a representative baseline from Boundless [31]. The training details are provided in the supplement.

Furthermore, to understand the impact of having the latent representation explicitly disentangled, we construct an uni-modal VQ-I2I variant as an additional baseline (denoted as uni-VQ-I2I). Specifically, in such uni-VQ-I2I baseline, we assume that the domain-specific style information is implicitly modeled by the generators $\{G_X, G_Y\}$, thus the domain-specific style features are discarded. Please refer to our supplementary materials for more details.

4.1   Qualitative Evaluation

I2I Translation on Unpaired and Paired Data. The proposed VQ-I2I synthesizes high-quality images on both the shape-invariant (winter-to-summer) and shape-variant (dog-to-cat, photo-to-portrait) datasets, as shown in Figure 3(a), where it achieves comparable or even better quality in comparison to other representative I2I methods. Moreover, the results of uni-modal VQ-I2I variant (denoted as uni-VQ-I2I), which excludes the disentanglement between content and style information are also provided, where we are able to observe that uni-VQ-I2I encounters the problem of texture inconsistency (*e.g.* there exists different styles in the cat's face on the first row of Figure 3(a)). The comparison between

our VQ-I2I and the uni-VQ-I2I variant reveals that the characteristic of disentanglement enables both content stability and style diversity, where we provide further explorations on uni-VQ-I2I in the supplementary materials. On the other hand, given pairs of semantic segmentation maps and corresponding images as training data, our proposed scheme produces appealing images that correspond to the input segmentation map (cf. Figure 3(b)). These results validate that our VQ-I2I approach can understand the semantic meaning of labels and synthesize correct instances, such as buildings and vehicles.

Multimodal Translation. Our VQ-I2I framework can also perform style-guided translation that produces diverse (multimodal) I2I results. Since vector-quantized content representation encodes the domain-invariant information while the style features carry the style information, we re-combine the same content with various styles to achieve diverse translations. The results are shown in the left portion of Figure 4. In addition to the inter-domain I2I, our method can also perform *intra*-domain I2I (as shown in the column labelled as "intra-domain" of Figure 4, in which we combine the content and style extracted from two images of the same domain), although we do not explicitly involve intra-domain I2I during the training stage.

Diverse Image Extension and Completion. The auto-regressive procedure built upon the content representation of VQ-I2I enables image extension. Specifically, as the content indices on the extended regions are drawn from the conditional distribution predicted by the transformer model, together with the style features being disentangled from content, the resultant extension produced by our VQ-I2I includes the diversity of both content and style (cf. Figure 5). It is worth noting that the extension results show that VQ-I2I generators would adjust the original image slightly to make the overall appearance of image extension more harmonious. Similar to image extension, our VQ-I2I is able to realize the image completion. We conduct the experiments of image completion on AFHQ [4] and Yosemite [34] dataset and provide some example results in the right portion of Figure 4, where only a quarter of an image is given as the input. Again, our auto-regressive model and the disentanglement designs are capable of generating diverse content and supporting style variability via combining the translation (*e.g.* inter-domain I2I in the rightmost column of Figure 4).

4.2   Quantitative Evaluation

We use the Fréchet inception distance (FID) [9] score and natural image quality evaluator (NIQE) [23] to measure the quality of the generated results and compare our proposed method to the existing approaches. Lower FID and NIQE scores indicate better perceptual quality. Moreover, we conduct a user study using the manner of pairwise comparison (*i.e.* our VQ-I2I versus baselines, or the images produced by various methods against the real images).

Table 1. Quantitative Comparisons with Unpaired I2I Methods. We measure the FID and NIQE scores across various datasets. VQ-I2I performs comparably to the state-of-the-art methods on unpaired datasets, while enabling applications that cannot be done by these conventional I2I methods.

	FID			NIQE	
	dog→cat	winter→summer	photo→portrait	dog→cat	winter→summer
CycleGAN	76.89	65.71	104.96	**40.65**	53.28
DRIT	35.74	**60.53**	102.52	47.32	**32.76**
MUNIT	33.78	94.78	**94.42**	63.64	35.64
U-GAT-IT	**21.62**	73.89	104.93	59.84	57.44
CUT	22.79	70.41	102.65	48.95	37.29
uni-VQ-I2I	25.65	62.43	99.37	45.41	36.53
VQ-I2I	26.53	63.60	100.29	53.29	35.97

Table 2. Quantitative Comparisons on Applications of VQ-I2I. (a) We evaluate the performance (in terms of FID scores) of unconditional generation on Yosemite [34] dataset via sampling 100 images respectively generated by our VQ-I2I and the VQGAN [7]. (b) Given the input image of size 256×256, we extend it horizontally for 50% and 75% (128 and 192 pixels respectively) toward the right-hand size, where we evaluate the FID scores on the right most portion of size 256×256 of the resultant image (*i.e.* this portion will recover part of the original input image and the extended region). The results show that our model is comparable to the existing extension method while extending for a larger range.

	256×256 generation
VQGAN	127.84
VQ-I2I	127.31

	outpaint for 50%	outpaint for 75%
Boundless [31]	**68.00**	**88.95**
VQ-I2I	77.82	90.05

(a) Unconditional Generation.   (b) Image Extension.

FID and NIQE. We summarize the FID and NIQE evaluation of unpaired I2I translation in Table 1 and FID measurement of paired I2I translation in Table 3. For unconditional generation, we compute the FID scores on the synthesized image of size 256×256, as shown in Table 2(a). As for image outpainting/extension, we present the quantitative results in Table 2(b), where the FID scores for image extension are computed from the distribution distance between the Yosemite dataset [34] and the rightmost 256×256 pixels of the extended images.

We are able to see that our proposed method performs comparably against the state-of-the-art translation frameworks, generative approach (*i.e.* VQGAN [7]) and the extension baselines (*i.e.* Boundless [31] and InfinityGAN [19]). Please note that, the main goal of our VQ-I2I is not to achieve superior performance in translation, unconditional generation or extension, instead we aim to facilitate both translation and the unconditional distribution shared among domains in a unified novel framework as well as unleash various interesting applications which other existing works are hard to realize (as described in the next subsection).

User Preference. To better rate the realism of I2I translation and image extension results, we conduct a user study with the manner of pairwise comparison.
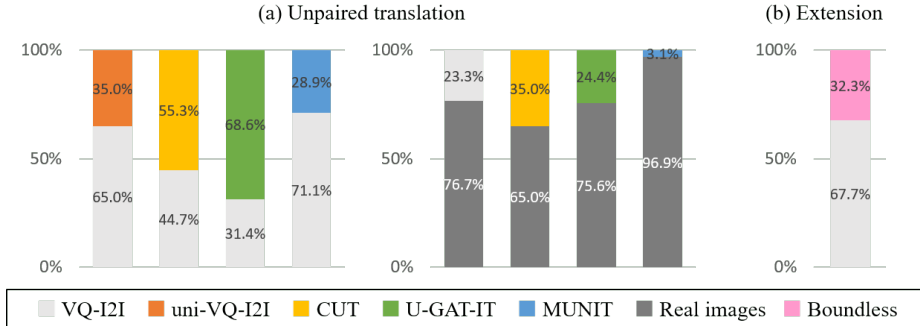
Fig. 6. User Preference Study. We conduct the user study (~180 participants) to compare VQ-I2I to different existing translation methods in (a), and the Boundless [31] method for the image extension task in (b).

Table 3. Quantitative Comparisons with Paired I2I Methods. We measure FID score for label→cityscapes translation on Cityscapes [5] Dataset.

	FID label→cityscapes
Pix2pix [13]	**51.73**
BicycleGAN [36]	93.13
VQ-I2I	74.03

Table 4. Ablation of Varying the Codebook Size and Dimensionality. We measure the FID score for summer→winter translation after 420 epochs of training on each model.

	Codebook Size	
	64	512
Dimensionality 64	96.94	96.71
Dimensionality 512	94.38	99.51

For I2I translation, each subject (in total ~180 participants) needs to answer the question "Which image is more realistic" given a pair of images (1) sampled from real images and the translated images generated from various I2I baselines or (2) respectively produced by our VQ-I2I and one of the baselines; while for extension, the comparison is conducted between our VQ-I2I and a baseline extension method (*i.e.* Boundless [31]). Figure 6 presents the results of the user study. The performance of VQ-I2I is comparable to those SOTA methods in I2I translation and image extension.

4.3 Applications

Unconditional Image Generation and Image Extension. VQ-I2I completes more applications that other existing pixel-level I2I models scarcely achieve, as we adapt the vector quantized representation to the disentangled domain-invariant content space. Combining generated or extended content codes with a replaceable style representation, VQ-I2I can be further utilized in two applications: unconditional image generation and image extension with flexible style modulation in different ways (i.e. the combination between generation/extension and intra- or inter-domain I2I), where we have demonstrate example results in
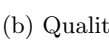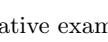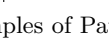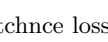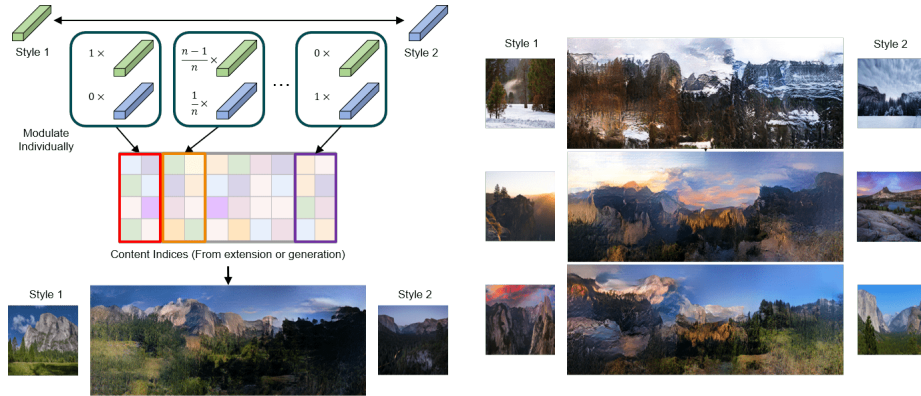
Fig. 7. Advanced Application of our VQ-I2I: Transitional Stylized Image Synthesis. Given two guided styles and the content map (produced from extension or unconditional generation), VQ-I2I is capable of synthesizing images with a smooth and gradually changing stylization effect via blending over two styles.

Figure 4, Figure 5, and Figure 1 (b)(c)(d). These applications afford to make image synthesis style-oriented, and there are more results provided in the supplementary materials.

Stylized Transitional Generation. In addition to single style modulation on the generated content, we can also perform multi-style transitional transfer via interpolating two styles to produce the style representation. As shown in Figure 7, We modulate different parts of the content map independently with different proportions by mixing the two styles, and merge all these modulated latents together to generate the transitional stylized output. In detail, for producing smooth and gradually changing effect of stylization, we partition the content map horizontally to 10 equal splits, where some example results are demonstrated in Figure 1(e) and Figure 7. More results with different number of splits are provided in the supplementary materials.

4.4   Further Investigation

Adding Patchwise Contrastive Loss. As our VQ-I2I framework does not include the cycle consistency as used in CycleGAN [34] or DRIT [17], there could exist a potential concern about being unable to well preserve the geometric information during I2I translation. To address this issue, here we experiment to adopt the patchwise contrastive loss [28], also named as PatchNCE loss, to enhance the content preservation during the training phase. As shown in Figure 8(a), the performance of using PatchNCE loss is more task-sensitive. Therefore, we consider it as an optional design choice, and use the content/style regression loss as the default design in our framework.

FID	VQ-I2I	+PatchNCE
dog→cat	**29.07**	80.72
winter→summer	**65.64**	71.17
photo→portrait	125.37	**114.04**

(a) Ablation study on adopting Patch-NCE Loss in our VQ-I2I model (performance in terms of FID scores).

(b) Qualitative examples of Patchnce loss.

Fig. 8. Ablation of Adding Patchwise Contrastive Loss (PatchNCE Loss). (a) We compute the FID scores with the same input content and style images on AFHQ, Yosemite, and Portrait datasets. The quantitative results reveal that PatchNCE loss makes a strong improvement on photo→portrait translation. (b) Given the input content and style images, the visual results show that PatchNCE loss is beneficial for our VQ-I2I model for preserving the content information on Portrait dataset [17].

Varying Codebook Size and Dimensionality. To observe the usage of codes in the codebook, we conduct additional ablation on Yosemite dataset [34] by varying the codebook size and the dimensionality of the codebook in VQ-I2I, and the FID scores for summer→winter translation is shown in Table 4. When setting the codebook size as 512 and dimensionality of the codebook as 512, our VQ-I2I model only uses around 35 codes. Besides, when shrinking both codebook size and dimensionality to 64, the codebook utilization grows up to 100%. However, from Table 4, the quantitative differences between different codebook size and dimensionality are imperceptible. Therefore, we suggest that training on Yosemite dataset [34] for a smaller codebook size and dimensionality still maintains its performance and reduces the memory usage of our VQ-I2I model.

5   Conclusion

In this paper, we introduce VQ-I2I, a novel image-to-image translation framework equipped with disentangled and discrete representations. In particular, our method learns a vector-quantized codebook for capturing the domain-invariant content information of input domains, in which such codebook enables the learning of the content distribution via an autoregressive model built upon the transformer network. Upon having comparable quantitative and qualitative performance at image-to-image translation with respect to several baselines, VQ-I2I is especially novel to have multifunctionality integrated into a unified framework, including image-to-image translation, unconditional generation, image extension, transitional stylization, and the combinations of the applications above.

References

1. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning (ICML) (2020)
2. Cheng, Y.C., Lee, H.Y., Sun, M., Yang, M.H.: Controllable image synthesis via segvae. In: European Conference on Computer Vision (ECCV) (2020)
3. Cheng, Y.C., Lin, C.H., Lee, H.Y., Ren, J., Tulyakov, S., Yang, M.H.: Inout: Diverse image outpainting via gan inversion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
4. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., Jiao, J.: Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
7. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
8. Han, L., Ren, J., Lee, H.Y., Barbieri, F., Olszewski, K., Minaee, S., Metaxas, D., Tulyakov, S.: Show me what and tell me how: Video synthesis via multimodal conditioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in Neural Information Processing Systems (NeurIPS) (2017)
10. Huang, H.P., Tseng, H.Y., Lee, H.Y., Huang, J.B.: Semantic view synthesis. In: European Conference on Computer Vision (ECCV) (2020)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: IEEE International Conference on Computer Vision (ICCV) (2017)
12. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: European Conference on Computer Vision (ECCV) (2018)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
14. Kim, J., Kim, M., Kang, H., Lee, K.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: International Conference on Learning Representations (ICLR) (2020)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2013)
16. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)

17. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: European Conference on Computer Vision (ECCV) (2018)
18. Lee, H.Y., Tseng, H.Y., Mao, Q., Huang, J.B., Lu, Y.D., Singh, M., Yang, M.H.: Drit++: Diverse image-to-image translation via disentangled representations. International Journal of Computer Vision (IJCV) (2020)
19. Lin, C.H., Lee, H.Y., Cheng, Y.C., Tulyakov, S., Yang, M.H.: Infinitygan: Towards infinite-pixel image synthesis. In: International Conference on Learning Representations (ICLR) (2021)
20. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
21. Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
22. Mao, Q., Tseng, H.Y., Lee, H.Y., Huang, J.B., Ma, S., Yang, M.H.: Continuous and diverse image-to-image translation via signed attribute vectors. International Journal of Computer Vision (IJCV) (2022)
23. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters (2012)
24. Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K.: Image to image translation for domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
25. van den Oord, A., Kalchbrenner, N.: Pixel rnn. In: International Conference on Machine Learning (ICML) (2016)
26. Oord, A.v.d., Kalchbrenner, N., Vinyals, O., Espeholt, L., Graves, A., Kavukcuoglu, K.: Conditional image generation with pixelcnn decoders. In: Advances in Neural Information Processing Systems (NeurIPS) (2016)
27. Oord, A.v.d., Vinyals, O., Kavukcuoglu, K.: Neural discrete representation learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
28. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision (ECCV) (2020)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
30. Razavi, A., van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
31. Teterwak, P., Sarna, A., Krishnan, D., Maschinot, A., Belanger, D., Liu, C., Freeman, W.T.: Boundless: Generative adversarial networks for image extension. In: IEEE International Conference on Computer Vision (ICCV) (2019)
32. Tseng, H.Y., Lee, H.Y., Jiang, L., Yang, M.H., Yang, W.: Retrievegan: Image synthesis via differentiable patch retrieval. In: European Conference on Computer Vision (ECCV) (2020)
33. Zhang, Z., Ma, J., Zhou, C., Men, R., Li, Z., Ding, M., Tang, J., Zhou, J., Yang, H.: Ufc-bert: Unifying multi-modal controls for conditional image synthesis (2021)
34. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)

35. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Multimodal image-to-image translation by enforcing bi-cycle consistency. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
36. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)