Taming Vector-Wise Quantization for Wide-Range Image Blending with Smooth Transition

Zeyu Wang
wangzeyu2020@zju.edu.cn
Zhejiang University
Hangzhou, China

Haibin Shen
shen_hb@zju.edu.cn
Zhejiang University
Hangzhou, China

Kejie Huang*
huangkejie@zju.edu.cn
Zhejiang University
Hangzhou, China

ABSTRACT

Wide-range image blending is a novel image processing technique that merges two different images into a panorama with a transition region. Conventional image inpainting and outpainting methods have been used to complete this task, but always create significant distorted and blurry structures. The State-Of-The-Art (SOTA) method uses a U-Net-like model with a feature prediction module for content inference. However, it fails to generate panoramas with smooth transitions and visual realness, particularly when the input images have distinct scenery features. It indicates that the predicted features may deviate from the natural latent distribution of authentic images. In this paper, we propose an effective deep-learning model that integrates vector-wise quantization for feature prediction. This approach searches for the most-like latent features from a discrete codebook, resulting in high-quality wide-range image blending. In addition, we propose to use the global-local discriminator for adversarial training to improve the predicted content quality and smooth the transition. Our experiments demonstrate that our method generates visually appealing panoramic images and outperforms baseline approaches on the Scenery6000 dataset.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

Wide-range image blending, vector-wise quantization, codebook, global-local discriminator.

ACM Reference Format:
Zeyu Wang, Haibin Shen, and Kejie Huang. 2023. Taming Vector-Wise Quantization for Wide-Range Image Blending with Smooth Transition. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3607541.3616809

*Corresponding author.

McGE '23, October 29, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0278-5/23/10…$15.00
https://doi.org/10.1145/3607541.3616809

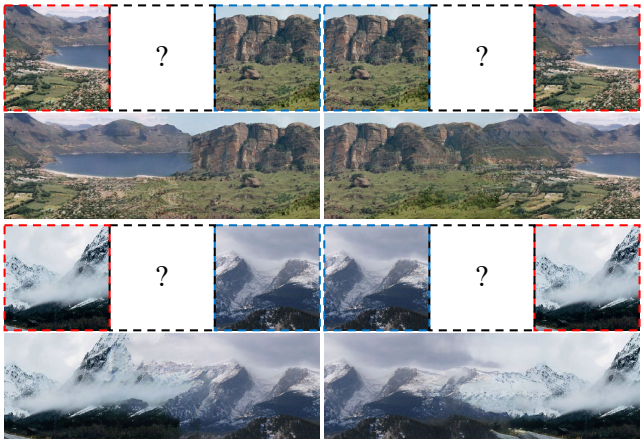
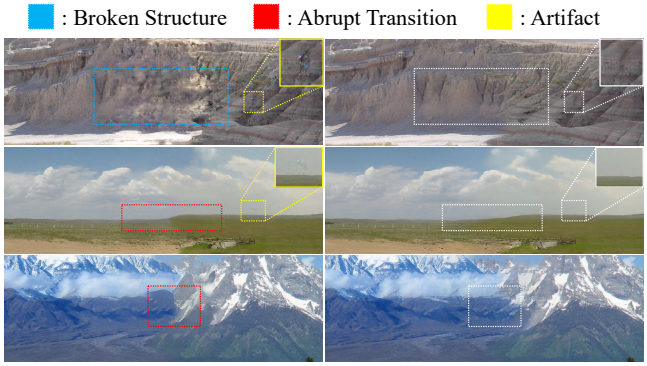


Figure 1: Examples of wide-range image blending with opposite input combinations. The resultant panoramas are generated by our proposed method.

1 INTRODUCTION

Wide-range image blending, a novel Computer-Vision (CV) problem scenario [16], has shown great potential in digital image processing. Its objective is to seamlessly merge two distinct input images into a panorama with an intermediate transition region. This technique is a creative method for image generation and composition, capable of generating 360° cyclic panoramas and long-range photos. While wide-range image blending shares similarities with image inpainting [1, 2, 18, 21] and image outpainting [7, 8, 13, 27], their task requirements are totally different. In image inpainting, the regions to be filled are usually narrow and irregular. In image outpainting, the contents to be generated are extended from the input images. In contrast, wide-range image blending requires the predicted intermediate content to be visually realistic and act as a smooth transition region, as illustrated in Figure 1.

In practice, image inpainting and outpainting methods have been utilized in wide-range image blending. In image inpainting methods, the intermediate transition region is treated as a missing area to be filled. In contrast, image outpainting methods extend the two input images separately and blend the extended contents into the transition region using an image blending model (*e.g.*, GP-GAN [26]). However, due to task incompatibility, both methods often create distorted and blurry structures, resulting in noticeable artifacts in the generated panoramic images. To address these issues, Lu *et al.* [16] proposed a U-Net-like [20] deep model, which is currently the State-Of-The-Art (SOTA) method for wide-range image



(a) SOTA Baseline (b) Our Method

Figure 2: Example inference results of (a) the SOTA baseline method and (b) our method.

blending. Their model uses a feature prediction module that infers the intermediate feature from the two-side features, connecting the encoder and decoder. Compared to previous methods, this approach significantly improves the quality of generated panoramic images. However, in some cases, the generated panoramas still contain abrupt transitions, broken structures, and noticeable artifacts, especially when the input images have distinct scenery features, such as different weather, as shown in Figure 2 (a). From our analysis, the above problems result from two drawbacks. (*i*) The predicted intermediate features may deviate from the natural latent distribution of authentic images. (*ii*) The network focuses primarily on the visual realness of predicted contents, with less attention paid to the global smoothness of the panoramas.

In this paper, we propose a novel deep-learning-based neural network for high-quality wide-range image blending. Specifically, we implement vector-wise quantization with a latent-discrete codebook for feature prediction. This approach prompts the predicted features to fit the latent distribution of authentic images, addressing the issue of deviation from natural latent distribution. Additionally, we design a global-local discriminator for adversarial training, focusing on both global smoothness and local visual realness. Our method solves the abovementioned problems and significantly improves the quality of generated panoramas, as shown in Figure 2 (b). Experimental results on the benchmark Scenery6000 dataset [27] show that our method outperforms the current state-of-the-art method and other baselines. Our implementation code will be online available until the publication of the paper.

2 RELATED WORK

2.1 Image Inpainting

Image inpainting, also known as image completion or image reconstruction, refers to filling the missing regions of images. This technique has several applications, including damaged image repair, object removal, and image-based rendering. Traditional image inpainting methods include diffusion-based methods [1, 2, 12], which use partial differential equations to propagate information from the surrounding regions to the hole region, and patch-based methods [5, 9, 21], which fill the hole region using patches extracted from the surrounding regions. While these methods have achieved promising results, they have several limitations, including difficulty in handling large holes and generating realistic textures. Recent works have employed deep neural networks for direct end-to-end inpainting [11, 14, 15, 18, 19], mostly based on convolution or Transformer methods. At present, image inpainting usually aims at filling irregular holes in images, and some works target pluralistic inpainting, which aims to generate diverse and plausible predictions for the missing regions [24, 33].

2.2 Image Outpainting

Image outpainting is an image processing technique that aims to generate new visual content beyond the original boundaries of an image. Different from image inpainting, the primary objective of image outpainting is to create extended content that is visually realistic and consistent with the input image. At present, this technique is predominantly based on deep neural networks, such as SpiralNet [8], U-Transformer [7], and others [4, 13, 27]. The major advantage of image outpainting is its ability to generate long-range or large-scale images. This capability has a wide range of potential applications in areas such as computer graphics and video processing.

3 METHODOLOGY

Given two input images I_l and I_r, The goal of high-quality wide-range image blending is to create a panorama $\widetilde{I}$ that appears visually realistic and has smooth transitions between them. To achieve this, we propose a novel neural network that employs vector-wise quantization for feature prediction. Meanwhile, we incorporate a global-local discriminator D_{gl} to enhance the adversarial training. Figure 3 illustrates our proposed model.

3.1 Panorama Generation

3.1.1 Network Pipeline. Our network, shown in Figure 3 (a), comprises four main components: the encoder $\mathcal{E}_\theta$, the decoder $\mathcal{G}_\phi$, the codebook $\mathcal{Z}$, and the Bidirectional Content Transfer (BCT) module $\mathcal{B}_\psi$. Specifically, $\mathcal{E}_\theta$ and $\mathcal{G}_\phi$ are composed of stacked convolutional residual blocks, connected with the attention-based Skip Horizontal Connection (SHC). The codebook $\mathcal{Z} = \{z_n\}_{n=1}^{N} \in \mathbb{R}^{N\times d}$ contains N discrete latent vectors, each with dimension d. Lastly, the Bidirectional Content Transfer (BCT) module $\mathcal{B}_\psi$ employs a bidirectional Long-Short Term Memory (LSTM) to perform feature prediction. The detailed structures of $\mathcal{E}_\theta$ and $\mathcal{G}_\phi$ can be found in the Supplementary section.

3.1.2 Latent Feature Encoding. Based on the two-side input images I_l and I_r, the encoder $\mathcal{E}_\theta$ encodes them into the latent features f_l and f_r, respectively. An additional linear projection layer $\mathbf{p}_{in}$ compresses them to z_l and z_r, which will be adopted for feature prediction.

$$f_l = \mathcal{E}_\theta(I_l), \quad f_r = \mathcal{E}_\theta(I_r) \tag{1}$$

$$z_l = \mathbf{p}_{in}(f_l), \quad z_r = \mathbf{p}_{in}(f_r) \tag{2}$$

where $I_l/I_r \in \mathbb{R}^{H\times W\times 3}$, $f_l/f_r \in \mathbb{R}^{h\times w\times D}$, and $z_l/z_r \in \mathbb{R}^{h\times w\times d}$. Since the encoder $\mathcal{E}_\theta$ has m down-sampling steps, the resultant $h = H/2^m$ and $w = W/2^m$. $\mathbf{p}_{in}$ denotes the projection layer.

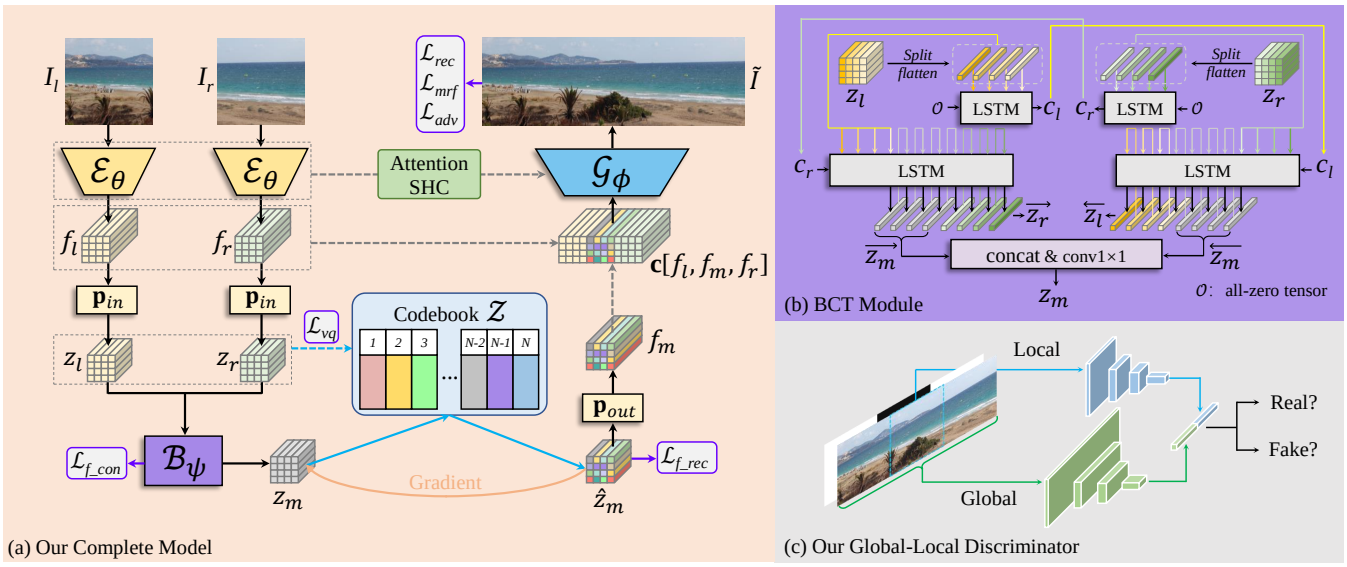


Figure 3: Illustration of our proposed model. (a) The pipeline of our complete model. (b) The Bidirectional Content Transfer (BCT) module. (c) Our proposed global-local discriminator ($\mathcal{D}_{gl}$).

3.1.3 Bidirectional Feature Prediction. For the panorama generation, the BCT module $\mathcal{B}_\psi$ is implemented to infer the intermediate latent feature z_m from z_l and z_r.

$$z_m = \mathcal{B}_\psi(z_l, z_r) \in \mathbb{R}^{h\times w\times d} \tag{3}$$

The structure of $\mathcal{B}_\psi$ is shown in Figure 3 (b). Based on the bidirectional LSTM, it predicts z_m along the two opposite directions, *i.e.*, from z_l to z_r and from z_r to z_l. The prediction is divided into two steps. The first step produces the temporary variables c_l and c_r based on z_l and z_r, respectively. The second step produces $\overrightarrow{z_m}$ and $\overrightarrow{z_r}$ from z_l and c_r, and produces $\overleftarrow{z_m}$ and $\overleftarrow{z_l}$ from z_r and c_l. For the feature consistency, $\overleftarrow{z_l}$ and $\overrightarrow{z_r}$ are expected to be identical to z_l and z_r, respectively. Then, the bidirectional variables $\overrightarrow{z_m}$ and $\overleftarrow{z_m}$ are concatenated and projected to our desired z_m. The design motivation of $\mathcal{B}_\psi$ can be found in [16], which is not the main focus of our work.

3.1.4 Vector-Wise Quantization. As mentioned in Introduction, the predicted intermediate feature z_m might deviate from the latent distribution of authentic images, resulting in abrupt transitions and broken structures in the generated panorama. To address this issue, we incorporate a latent-discrete codebook $\mathcal{Z}$, which aims to adjust z_m to fit the desired distribution.

In our proposed model, z_l and z_r represent the latent features of I_l and I_r, respectively. Therefore, to prompt $\mathcal{Z}$ to fit the latent distribution of authentic images, we encourage the discrete vectors in $\mathcal{Z}$ to learn from z_l and z_r. We train the codebook by minimizing the vector-quantization loss $\mathcal{L}_{vq}$, which measures the quantization distance. Crucially, after pre-training the codebook $\mathcal{Z}$, we apply the vector-wise quantization $\mathbf{q}(\cdot)$ on z_m, which generates $\hat{z}_m$. The vector-wise quantization involves searching for the closest code z_n from the codebook $\mathcal{Z}$ for the corresponding spatial latent feature z_m^{ij}.

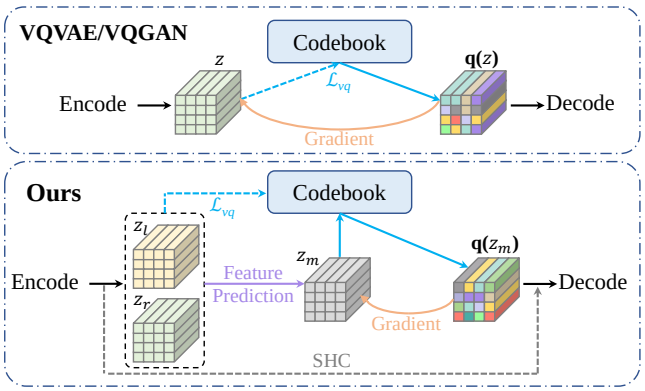


Figure 4: The comparison between the vector quantization processes of VQVAE/VQGAN and our method. Our design is tailored to the task of wide-range image blending.

$$\hat{z}_m = \mathbf{q}(z_m) := \left(\underset{z_n\in\mathcal{Z}}{\mathbf{argmin}} \|z_m^{ij} - z_n\|\right) \in \mathbb{R}^{h\times w\times d} \tag{4}$$

where $z_m/\hat{z}_m \in \mathbb{R}^{h\times w\times d}$, $z_m^{ij}/z_n \in \mathbb{R}^d$, $i \in \{0, 1, \ldots, h-1\}$, and $j \in \{0, 1, \ldots, w-1\}$. The quantization process $\mathbf{q}(\cdot)$ searches for the closest code z_n of the corresponding spatial latent feature z_m^{ij}, where z_n is selected from the codebook $\mathcal{Z}$.

It is worth noting that both VQVAE [23] and VQGAN [6] use a quantization process that encourages the network and codebook to learn to fit each other. In contrast, our proposed method trains the discrete vectors in $\mathcal{Z}$ to fit the distribution of z_l and z_r, with the quantization process only applied to z_m, as depicted in Figure 4. This design is tailored to the task of wide-range image blending.

3.1.5 Latent Feature Decoding. Before the decoding process, $\hat{z}_m$ is projected to the feature $f_m \in \mathbb{R}^{h\times w\times D}$. Conditioned on f_l, f_m, and f_r, we concatenate them along the horizontal direction to produce $\mathbf{c}[f_l, f_m, f_r] \in \mathbb{R}^{h\times 3w\times D}$. The concatenated feature is then decoded by $\mathcal{G}_\phi$ to generate the resultant panorama $\widetilde{I}$.

$$f_m = \mathbf{p}_{out}(\hat{z}_m) \tag{5}$$

$$\widetilde{I} = \mathcal{G}_\phi(\mathbf{c}[f_l, f_m, f_r]) \in \mathbb{R}^{H\times 3W\times 3} \tag{6}$$

where $\mathbf{c}$ denotes the concatenation process. The encoder $\mathcal{E}_\theta$ and decoder $\mathcal{G}_\phi$ are connected by the skip horizontal connection to form the U-Net-like structure.

3.2 Global-Local Adversarial Training

To enhance the realness and smoothness of the panorama transitions, we propose a novel global-local discriminator ($\mathcal{D}_{gl}$) for adversarial training, as illustrated in Figure 3 (c). It is inspired by the discriminator design in [11] for image inpainting.

In detail, $\mathcal{D}_{gl}$ consists of two components, *i.e.*, the global discriminator and the local discriminator. During adversarial training, the predicted panorama $\widetilde{I}$ is split into three parts, namely $\widetilde{I}_l$, $\widetilde{I}_m$, and $\widetilde{I}_r$. The global discriminator assesses the authenticity of the entire panorama $\widetilde{I}$, while the local discriminator focuses solely on $\widetilde{I}_m$. Additionally, we concatenate a $0-1$ mask to the panorama, where 0 and 1 represent the intermediate and two-side regions, respectively. The global and local discriminators employ the same network architecture with stacked convolutional layers, except for the last fully connected layers, which output vectors of the same dimension. Finally, the vectors are concatenated and linearly projected to a constant value that represents the authenticity of the generated panorama.

3.3 Training Process

During the training process, we take the images from the training set and crop them to a fixed shape. The ground-truth panorama $I \in \mathbb{R}^{H\times 3W\times 3}$ is split into I_l, I_m, and I_r from left to right along the horizontal direction. Conditioned on I_l and I_r, our network generates the resultant panoramic image $\widetilde{I}$. The training process aims to enforce the predicted $\widetilde{I}$ to be as close to the ground-truth I as possible. Since I_l and I_r are available during the training process, the main focus is the prediction of the intermediate content I_m, which is formulated as:

$$\mathbf{min}\, \mathbb{E}_{I_l, I_m, I_r \sim p(I)}[-\log p(I_m | I_l, I_r)] \tag{7}$$

Unlike the SOTA method [16], we do not adopt the fine-tuning stage, which utilizes additional training samples of I_l and I_r obtained from different images. During our extensive experiments, the additional fine-tuning stage does not bring quantitative or qualitative improvement and sometimes causes the instable training.

3.4 Loss Functions

The total loss function for training our network is formulated as:

$$\begin{aligned}\mathcal{L}(\mathcal{E}_\theta, \mathcal{G}_\phi, \mathcal{B}_\psi, \mathcal{Z}) = \lambda_1 \cdot \mathcal{L}_{rec} + \lambda_2 \cdot \mathcal{L}_{mrf} + \lambda_3 \cdot \mathcal{L}_{vq} \\ + \lambda_4 \cdot \mathcal{L}_{f\_rec} + \lambda_5 \cdot \mathcal{L}_{f\_con} + \lambda_6 \cdot \mathcal{L}_{adv}\end{aligned} \tag{8}$$

In detail, the hyperparameters are set as: $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 1$, $\lambda_4 = 1$, $\lambda_5 = 1$, $\lambda_6 = 0.0018$. The detailed loss terms will be illustrated in the following subsections.

3.4.1 Reconstruction Loss. Firstly, $\mathcal{L}_2$ is adopted as the reconstruction loss $\mathcal{L}_{rec}$. Notably, $\mathcal{L}_{rec}$ on the intermediate region is computed by involving a weighted mask M.

$$\mathcal{L}_{rec} = \|M \odot (\widetilde{I}_m - I_m)\|_2 + \|\widetilde{I}_l - I_l\|_2 + \|\widetilde{I}_r - I_r\|_2 \tag{9}$$

$$M(d) = \exp(-\frac{1}{2}(\frac{d}{\sigma})^2) + \exp(-\frac{1}{2}(\frac{d-W}{\sigma})^2) \tag{10}$$

where $\sigma = W/4$, and $d \in [0, W]$ denotes the horizontal position of I_m from left to right. With the weighted mask M, the pixels closer to the middle are less penalized for the panorama generation with more flexibility.

3.4.2 IDMRF Loss. Secondly, the Implicit Diversified Markov Random Fields (IDMRF) is adopted to improve the image texture consistency, which has been used in some prior works of image editing [17, 25]. The computation of IDMRF is based on the pre-trained VGG-19 network [22], which measures the distance between the feature maps of $\widetilde{I}_m$ and I_m.

$$\mathcal{L}_{mrf} = \text{IDMRF}(\widetilde{I}_m, I_m) \tag{11}$$

3.4.3 Vector-Quantization Loss. Thirdly, the Vector-Quantization (VQ) loss $\mathcal{L}_{vq}$ is employed to train the codebook $\mathcal{Z}$. The quantization distance is measured on z_l and z_r, but not on z_m. The minimized quantization distance helps the discrete vectors in $\mathcal{Z}$ fit the natural latent distribution of authentic images.

$$\begin{aligned}\mathcal{L}_{vq} =& \|\text{sg}[z_l] - \mathbf{q}(z_l)\|_2^2 + \beta\|\text{sg}[\mathbf{q}(z_l)] - z_l\|_2^2 \\ &+ \|\text{sg}[z_r] - \mathbf{q}(z_r)\|_2^2 + \beta\|\text{sg}[\mathbf{q}(z_r)] - z_r\|_2^2\end{aligned} \tag{12}$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operation, and β denotes the weighting factor, which is set to 0.25 in this work. It is worth noting that the quantized features $\mathbf{q}(z_l)$ and $\mathbf{q}(z_r)$ are calculated for computing $\mathcal{L}_{vq}$ during the training process, but they are not used during the forward process for inference in the network.

3.4.4 Feature Reconstruction Loss. Meanwhile, the feature reconstruction loss $\mathcal{L}_{f\_rec}$ is employed to measure the distance between $\hat{z}_m$ and the encoded feature of I_m, which smooths the training of $\mathcal{E}_\theta$ and $\mathcal{B}_\psi$. It is formulated as:

$$\mathcal{L}_{f\_rec} = \|\hat{z}_m - \mathbf{p}_{in}(\mathcal{E}_\theta(I_m))\|_2 \tag{13}$$

3.4.5 Feature Consistency Loss. In addition, the feature consistency loss $\mathcal{L}_{f\_con}$ is employed for training $\mathcal{B}_\psi$, which is vital for the bidirectional feature prediction. It is formulated as:

$$\mathcal{L}_{f\_con} = \|z_r - \overrightarrow{z_r}\|_2 + \|\overrightarrow{z_m} - \overleftarrow{z_m}\|_2 + \|z_l - \overleftarrow{z_l}\|_2 \tag{14}$$

3.4.6 Adversarial Loss. Finally, the adversarial loss of Relativistic Average Least-Square GAN (RaLSGAN) [17] is adopted with our global-local discriminator $\mathcal{D}_{gl}$, which is formulated as:

$$\mathcal{L}_{adv} = \sum \mathcal{D}_{gl}(\widetilde{I}, I)^2 \tag{15}$$

Simultaneously, the discriminator loss $\mathcal{L}_{adv}^{gl}$ penalizes $\mathcal{D}_{gl}$, which is formulated as below:

$$\mathcal{L}_{adv}^{gl}(\mathcal{D}_{gl}) = \sum [\mathcal{D}_{gl}(I, \widetilde{I}) - 1]^2 + [\mathcal{D}_{gl}(\widetilde{I}, I) + 1]^2 \tag{16}$$

Figure 5: Qualitative results of all baseline methods and our proposed method.

4 EXPERIMENTS

4.1 Experimental Details

4.1.1 Baselines. The baseline models for comparison include conventional image inpainting and outpainting methods, and the SOTA wide-range image blending method. Image inpainting methods include: **CA** [29], **Pen-Net** [30], **StrucFlow** [19], **HiFill** [28], and **ProFill** [31]. Image outpainting methods include: **SRN** [25] and **Yang *et al.*** [27]. The SOTA wide-range image blending method is: **Lu *et al.*** [16].

4.1.2 Dataset. We adopt the benchmark Scenery6000 dataset [27] in our experiments. The dataset consists of 6040 wide-range natural scenery images, where 5040 images are used for training and 1000 images are used for testing.

4.1.3 Metrics. We report the conventional metrics in image generation tasks for quantitative comparison: Fréchet Inception Distance (FID) [10] and Kernel Inception Distance (KID) [3]. The quantitative results of the baseline models are taken from [16]. In addition, we adopt Learned Perceptual Image Patch Similarity (LPIPS) [32] for further evaluation in our ablation study. All of these metrics are the lower the better.

Table 1: Quantitative results of all baselines and our method. ↓ means lower is better. Since the two-side regions only need to be reconstructed, the central 256×512 areas from the 256×768 panoramas are cropped for evaluation (as well for the baselines).

Method		FID↓	KID↓
Inpainting	CA [29]	91.87	0.0745
	Pen-Net [30]	159.70	0.1151
	StrucFlow [19]	138.13	0.2168
	HiFill [28]	139.39	0.1230
	ProFill [31]	46.53	0.0326
Outpainting	SRN [25]	70.94	0.0392
	Yang *et al.* [27]	82.69	0.0446
W-Blending	Lu *et al.* [16]	36.13	0.0116
	Ours	**32.52**	**0.0042**

4.1.4 Implementation. We adopt the Adam optimizer for training our proposed method. The total number of training epochs is 500. The learning rate is initially set to $1e^{-3}$ for the first 300 epochs, after which it is adjusted to $1e^{-4}$ for the remaining 200 epochs. During training, we randomly crop the training images to the size of 256×768. For fair evaluation, we crop the testing images from the center to 256×768.

4.2 Qualitative Comparison

Figure 5 shows the qualitative comparison between all baselines and our method. Obviously, these inpainting and outpainting methods exhibit poor performance for wide-range image blending. Among these methods, CA, StrucFlow, Hi-Fill, and Yang *et al.* generate panoramas with distorted structures and abrupt transitions, while Pen-Net, ProFill, and SRN generate blurry transition contents. In contrast, Lu *et al.* and our method perform well and significantly outperform the other methods.

To provide a more extensive comparison of the SOTA baseline (Lu *et al.*) with our method, we present more qualitative results in Figure 6. As can be seen, our method tends to generate panoramic images of better quality. From the comparison, we can summarize several advantages of our method: (*i*) Our method rarely generates broken or blurry structures, dramatically improving the visual realness. (*ii*) Our method exhibits better transition smoothness, especially when input images have distinct scenery features. (*iii*) Our method can avoid generating repetitive structures, as seen in the third result of Figure 6.

Figure 6: Extensive qualitative comparison. (a) Input images. (b) The SOTA baseline (Lu *et al.*). (c) Our proposed method.

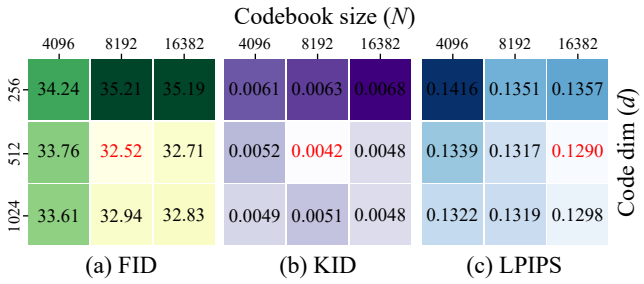


Figure 7: The quantitative results of our models with different codebook sizes (N) and code dimensions (d). The values marked in red represent the best results.

4.3 Quantitative Comparison

Table 1 shows the quantitative results of baselines and our method. Compared with these inpainting and outpainting methods, Lu *et al.* and our method prove that the specially designed methods exhibit better performance. Therefore, these previous methods are unsuitable for wide-range image blending due to task incompatibility.

More importantly, our method outperforms Lu *et al.* on all metrics. Firstly, our method achieves the FID score of 32.52, which is 3.61 lower than that of the SOTA baseline (Lu *et al.*). In addition, we achieve the KID score of 0.0042, which is only 36.2% of Lu *et al.* The improvement in FID and KID demonstrates that the vector-wise quantization reduces the latent distance between the generated images and the ground-truth images, which helps generate panoramas with more realness. Overall, the quantitative improvement illustrates the advantages of our proposed method.

Since the vector-wise quantization brings dramatic performance improvement, the codebook size N and code dimension d will influence the performance. We test models with different N and d, and compare the quantitative results in Figure 7. From the comparison, the code dimension d is preferably set to 512 for the best performance. The best FID and KID are achieved when the codebook size N equals 8192, while the best LPIPS is obtained when N equals 16382. Furthermore, we try to increase N and d, but no further performance improvement is achieved. Therefore, the codebook $\mathcal{Z} \in \mathbb{R}^{8192\times512}$ is adopted for our model.

4.4 Ablation Study

To illustrate the contribution of each loss objective and component of our method, we conduct the ablation study. In detail, we implement the variants in Table 2 with the same experimental settings and present the quantitative results. In addition, an example qualitative result for comparison is shown in Figure 8.

The ablation study shows that the lack of any loss objective or component will negatively influence the performance. Among these variants, the performance becomes the worst without $\mathcal{L}_{f\_con}$. The qualitative result shows that the corresponding variant performs poorly in the transition smoothness. On the contrary, the effect of $\mathcal{L}_{f\_rec}$ on the quantitative results is relatively tiny. It demonstrates that the vector-wise quantization has made $\hat{z}_m$ greatly fit the authentic distribution. The lack of $\mathcal{L}_{mrf}$ and the attention module will destroy the detailed textures. Evidently, without the codebook $\mathcal{Z}$ and $\mathcal{L}_{vq}$, the quantitative results are moderately influenced, and the qualitative result seems to exhibit worse details, demonstrating the effectiveness of vector-wise quantization. In addition, the lack of $\mathcal{D}_{gl}$ and $\mathcal{L}_{adv}$ dramatically affects the performance, where the

Table 2: Ablation Study: Quantitative comparison between all variants and our complete model. ↓ means lower is better. w/o denotes "without".

Variant		FID↓	KID↓	LPIPS↓
Objective	w/o $\mathcal{L}_{rec}$	35.56	0.0053	0.1399
	w/o $\mathcal{L}_{mrf}$	43.45	0.0097	0.1620
	w/o $\mathcal{L}_{f\_rec}$	32.61	0.0043	0.1319
	w/o $\mathcal{L}_{f\_con}$	79.49	0.0408	0.1791
Component	w/o attention	42.50	0.0102	0.1466
	w/o $\{\mathcal{Z}, \mathcal{L}_{vq}\}$	35.62	0.0063	0.1412
	w/o $\{\mathcal{D}_{gl}, \mathcal{L}_{adv}\}$	67.21	0.0327	0.1513
	$\mathcal{D}_{gl} \rightarrow \mathcal{D}_g$	34.46	0.0054	0.1343
Complete Model		**32.52**	**0.0042**	**0.1317**

Figure 8: Ablation Study: Qualitative comparison between all variants and our model. (a) w/o $\mathcal{L}_{rec}$. (b) w/o $\mathcal{L}_{mrf}$. (c) w/o $\mathcal{L}_{f\_rec}$. (d) w/o $\mathcal{L}_{f\_con}$. (e) w/o attention. (f) w/o $\{\mathcal{Z}, \mathcal{L}_{vq}\}$. (g) w/o $\{\mathcal{D}_{gl}, \mathcal{L}_{adv}\}$. (h) $\mathcal{D}_{gl} \rightarrow \mathcal{D}_g$. (i) Our Model.

corresponding qualitative result shows that the visual realness is extremely poor. Finally, when we replace $\mathcal{D}_{gl}$ with the conventional discriminator $\mathcal{D}_g$, the metrics become slightly worse, which illustrates the effectiveness of $\mathcal{D}_{gl}$.

4.5 Cyclic and Long-Range Panorama Generation

One application of our wide-range image blending method is the generation of images with 360° cyclic panoramas, as shown in Figure 9 (a). Given two different input images, it generates two composite

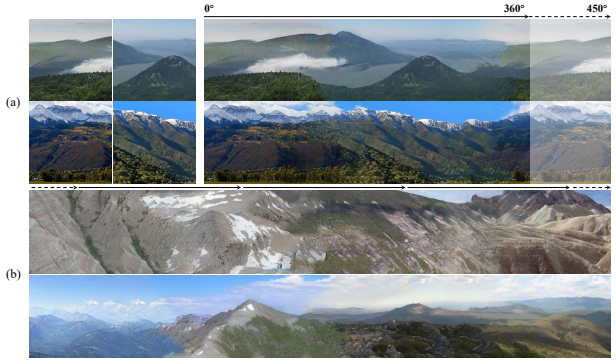


Figure 9: Cyclic and long-range panorama generation by our method. (a) Generation of 360° panoramic views from two input images. (b) Long-range image outpainting, generating the extended image with 2N-1 times the width from N input images.

images with opposite input orders (*i.e.* input 1→ input 2 and input 2→ input 1). The two composite images are then stitched together, generating the cyclic panoramas. It is valuable for mobile terminals when the cameras are limited in the field of view. Another application is that it can achieve wide-range image outpainting [27] in a unique way, as shown in Figure 9 (b). If we have N input images, we can arrange them horizontally in a row with N-1 spaces between them. Each space is inserted with the intermediate blended image generated from the neighboring input images. In this way, we can get an ultra-wide image with a width of 2N-1 times. The limitation lies in that, in some cases, it is difficult to obtain N different input images with similar scenery features, which will influence the quality of generated panoramas.

5 CONCLUSION

In this paper, we propose a novel deep-learning model with vector-wise quantization for wide-range image blending. Crucially, the latent-discrete codebook is implemented to represent the latent distribution of authentic images and prompt the predicted feature to fit it. In addition, the novel global-local discriminator is designed for adversarial training to improve visual realness and transition smoothness. Through the extensive experiments, the qualitative and quantitative results demonstrate that our method outperforms the SOTA and other baselines. The ablation study illustrates the significance of the loss objectives and proposed modules. Meanwhile, it also brings creative image-editing applications, including 360° cyclic panorama generation and long-range image outpainting. In the future, we would like to extend the task to flexible-range image blending, which can blend the input images into free-width panoramas.

ACKNOWLEDGMENTS

This research has been supported by National Natural Science Foundation of China (Grant No. 62274142) and Hangzhou Major Technology Innovation Project of Artificial Intelligence (Grant No. 2022AIZD0060).

REFERENCES

[1] Naoufal Amrani, Joan Serra-Sagristà, Pascal Peter, and Joachim Weickert. 2017. Diffusion-based inpainting for coding remote-sensing data. *IEEE Geoscience and Remote Sensing Letters* 14, 8 (2017), 1203–1207.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 417–424.

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401* (2018).

[4] Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. 2022. InOut: Diverse Image Outpainting via GAN Inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11431–11440.

[5] Maxime Daisy, David Tschumperlé, and Olivier Lézoray. 2013. A fast spatial patch blending algorithm for artefact reduction in pattern-based image inpainting. In *SIGGRAPH Asia 2013 Technical Briefs*. 1–4.

[6] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12873–12883.

[7] Penglei Gao, Xi Yang, Rui Zhang, Kaizhu Huang, and Yujie Geng. 2022. Generalised Image Outpainting with U-Transformer. *arXiv preprint arXiv:2201.11403* (2022).

[8] Dongsheng Guo, Hongzhi Liu, Haoru Zhao, Yunhao Cheng, Qingwei Song, Zhaorui Gu, Haiyong Zheng, and Bing Zheng. 2020. Spiral generative network for image extrapolation. In *European Conference on Computer Vision*. Springer, 701–717.

[9] Qiang Guo, Shanshan Gao, Xiaofeng Zhang, Yilong Yin, and Caiming Zhang. 2017. Patch-based image inpainting via two-stage low rank approximation. *IEEE transactions on visualization and computer graphics* 24, 6 (2017), 2023–2036.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–14.

[12] Haodong Li, Weiqi Luo, and Jiwu Huang. 2017. Localization of diffusion-based inpainting in digital images. *IEEE transactions on information forensics and security* 12, 12 (2017), 3050–3064.

[13] Han Lin, Maurice Pagnucco, and Yang Song. 2021. Edge guided progressively generative image outpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 806–815.

[14] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*. 85–100.

[15] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4170–4179.

[16] Chia-Ni Lu, Ya-Chu Chang, and Wei-Chen Chiu. 2021. Bridging the visual gap: Wide-range image blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 843–851.

[17] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

[18] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.

[19] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 181–190.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[21] Tijana Ružić and Aleksandra Pižurica. 2014. Context-aware patch-based image inpainting using Markov random field modeling. *IEEE transactions on image processing* 24, 1 (2014), 444–456.

[22] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2015).

[23] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[24] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4692–4701.

[25] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. 2019. Wide-context semantic image extrapolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1399–1408.

[26] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2019. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*. 2487–2495.

[27] Zongxin Yang, Jian Dong, Ping Liu, Yi Yang, and Shuicheng Yan. 2019. Very long natural scenery image prediction by outpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10561–10570.

[28] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. 2020. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7508–7517.

[29] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5505–5514.

[30] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. 2019. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1486–1494.

[31] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. 2020. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European conference on computer vision*. Springer, 1–17.

[32] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[33] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1438–1447.