A. Baseline Inpainting Methods More Details

EdgeConnect It consists of the edge generation network and the image completion network. The inputs of an edge generator are a mask, a gray image, and an edge map computed using Canny edge detector. The edge generator is trained to predict edges in the masked area of an edge map. This predicted edge map is then used as the prior information for the second stage. The image completion network uses a masked image to fill the unknown area with the generated edge map, which helps minimize blurriness by providing sketches for inpainting. In this method, the prior information and predicted edge map play a major role in reconstructing realistic structures with minimal blurriness. However, the results are highly dependent on the edge map, so if the quality of the generated edges is not good, results of the image completion will be unsatisfactory.

RFR-Net It has two main components: RFR module and Knowledge Consistent Attention (KCA) module. RFR consists of three modules. In the first module, *i.e.*, the area identification module, the partial convolution layer identifies the area by shrinking the mask in each recurrence. After passing through a partial convolution, the generated feature map becomes the input of the second module, *i.e.*, the feature reasoning module. In the feature reasoning module, there are simple encoding and decoding layers, which output a feature map having the same size as the input size. The output of the feature reasoning module becomes the input of the area identification by the skip connection. The two modules continue to work alternately and recurrently, generating output feature maps in each recurrence until the missing regions are filled. In the third module, the feature merging module, *i.e.*, the generated feature maps are merged by optionally averaging the values in each feature map if its locations have been filled. The RFR module processes in the feature map space whereas other progressive methods output the same representation as the input. This difference results in a much lighter model because the module reuses parameters and can be flexibly moved in the network. KCA makes the attention module consistent by adaptively combining attention scores from multiple recurrences. If attention scores of recurrences are calculated independently, it could be inconsistent. However, the KCA module adaptively combines the attention score of the previous recurrence with the attention score of the current recurrence using learnable parameters. Hence, it assists the inferential process of the RFR module by ensuring consistency between patch-swapping processes among recurrences. In this method, RFR-Net progressively strengthens the constraints for filling in large missing regions. The results of RFR-Net become semantically improved.

B. Additional Results

In this section, we show a comparison of the results for more images. For fair comparison, we randomly selected the images and compared them. Fig. 9 is a comparison of the results on the SUN dataset, and Fig. 10 and Fig. 11 are comparison of the results on the Beach dataset.

C. Code Descriptions

Our code is based on PyTorch version of Vision Transformer. We use ViT Generator to generate Hint images. We set the parameters as follows: image_size_h = 256, image_size_w = 128, patch_size_h = 16, patch_size_w = 16, dim = 1024, depth = 6, heads = 16, mlp_dim = 2048, dropout = 0.1, emb_dropout = 0.1.

Code is available at
`https://github.com/kdh4672/hgonet`

D. Detailed Description of Evaluation Metrics

We used reference IQA and No-reference IQA to diagnose the effect of the proposed method. Each evaluation metric is described in detail.

D.1. Fréchet Inception Distance (FID)

The FID compares the distribution of the generated images with the distribution of real images. The FID measures the distance between a generated distribution and real dataset distribution, as approximated by a Gaussian fit to samples projected into the feature space of a pretrained Inceptionv3 model. The FID is defined as follows:

$$\text{FID} := \|m - m_w\|_2^2 + Tr\left(C + C_w - 2(CC_w)^{1/2}\right) \quad (1)$$

where m and C are the mean and co-variances of the inception embeddings for real-data, and m_w and C_w are the mean and covariance matrices of the inception embeddings for the generated samples. The FID correlates well with image quality, and is capable of detecting mode collapse.

D.2. Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE)

The no-reference algorithm evaluates image quality using statistical characteristics of the input image. BRISQUE is an algorithm based on the idea that when a natural image is distorted, the statistics of the image pixels are also distorted. It extracts the pointwise statistics of local normalized luminance signals and measures image naturalness based on measured deviations from a natural image model. In general, it is known that all no-reference quality metrics are superior to fully referenced metrics in terms of consistency with human subjective quality scores.

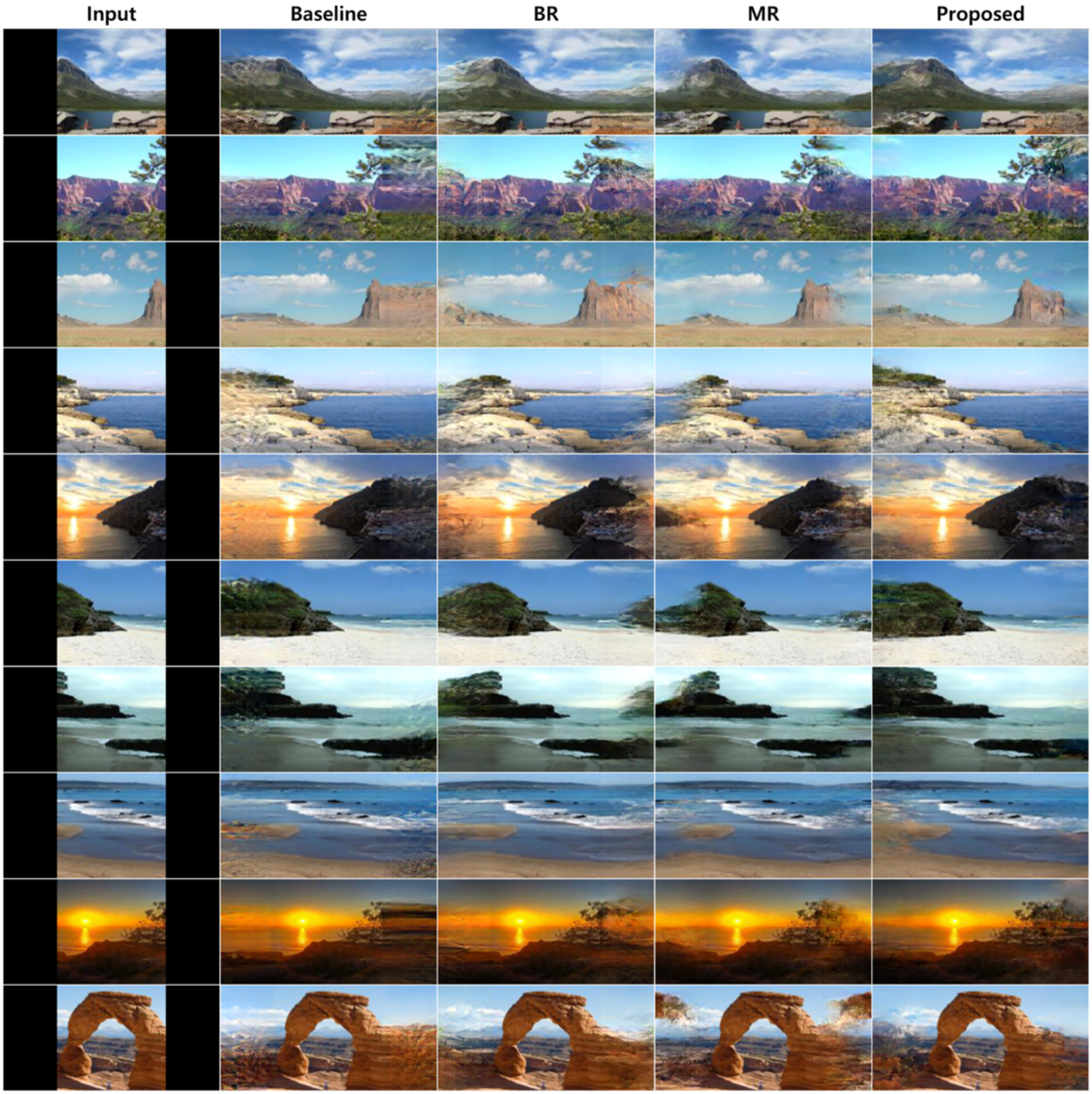


Figure 9. Qualitative results for conventional and proposed methods on SUN dataset.

E. Subjective Test using Mean Opinion Score

In all conducted MOS tests we have asked 30 human raters to assign a score from 1 (Bad) to 5 (excellent) to the outpainted results. In total 900 ratings were obtained, where each rater rated 30 images. We conducted a survey using the following survey form. (Fig. 12).

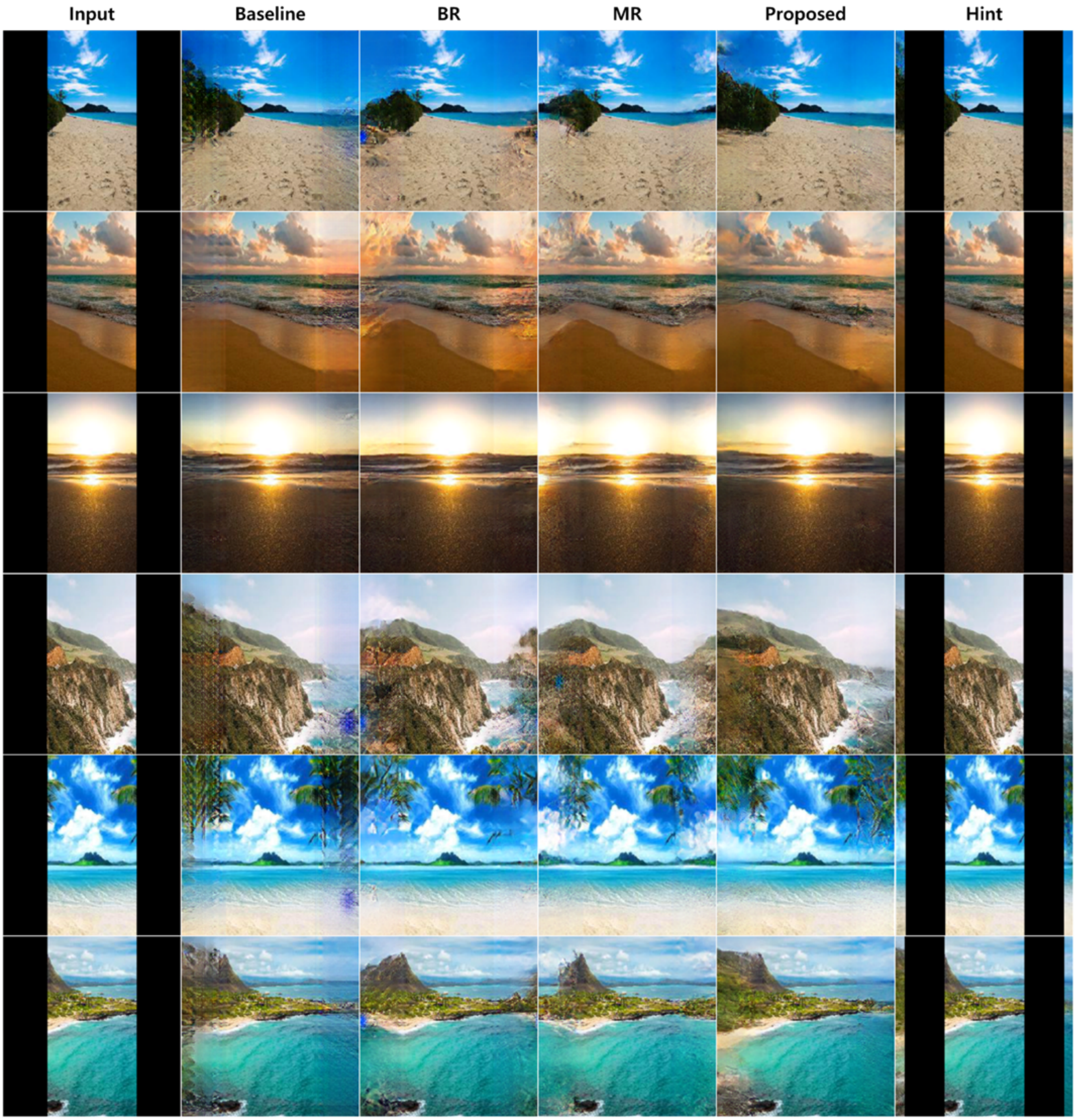


Figure 10. Qualitative results for conventional and proposed methods on Beach dataset.

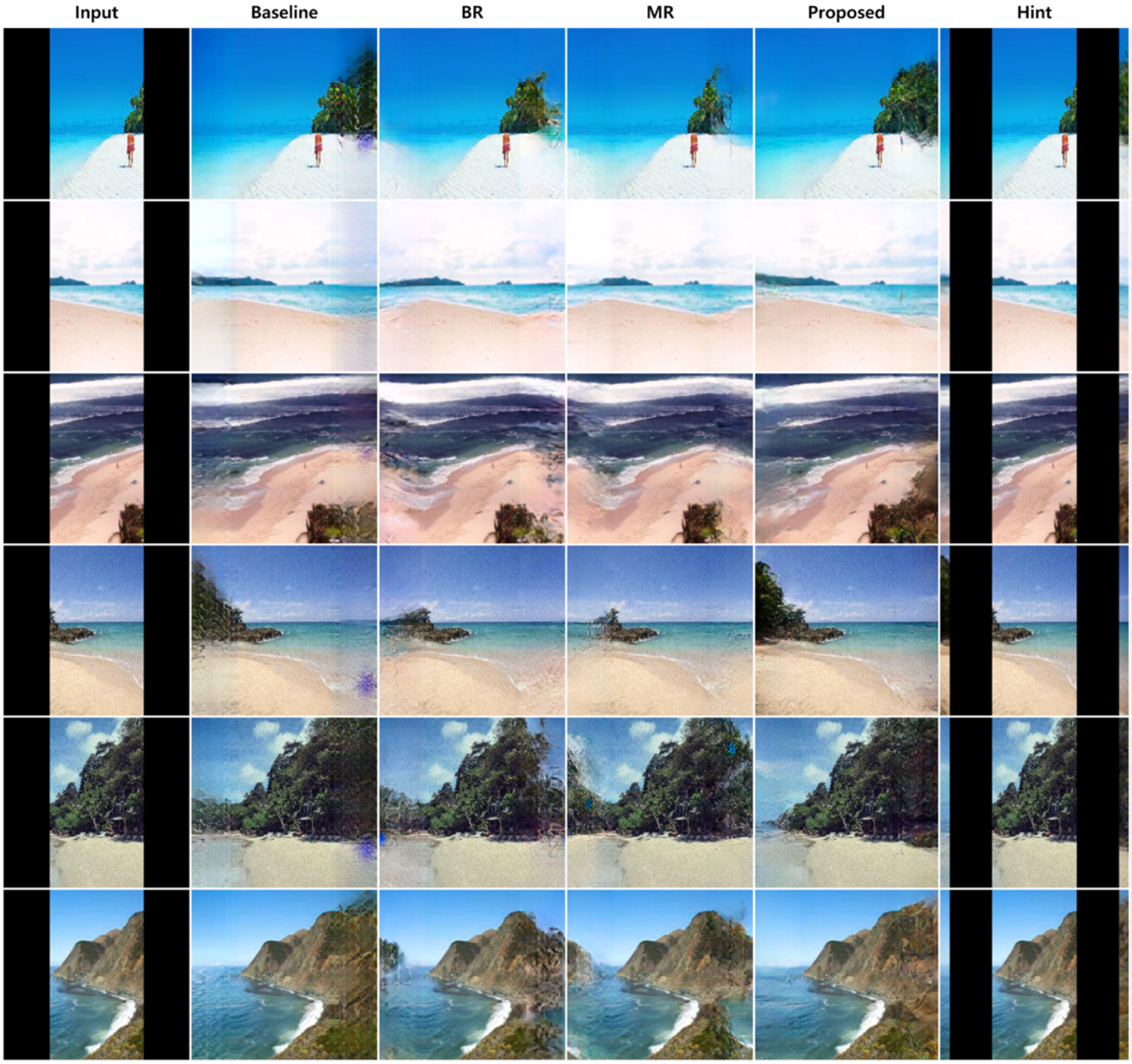
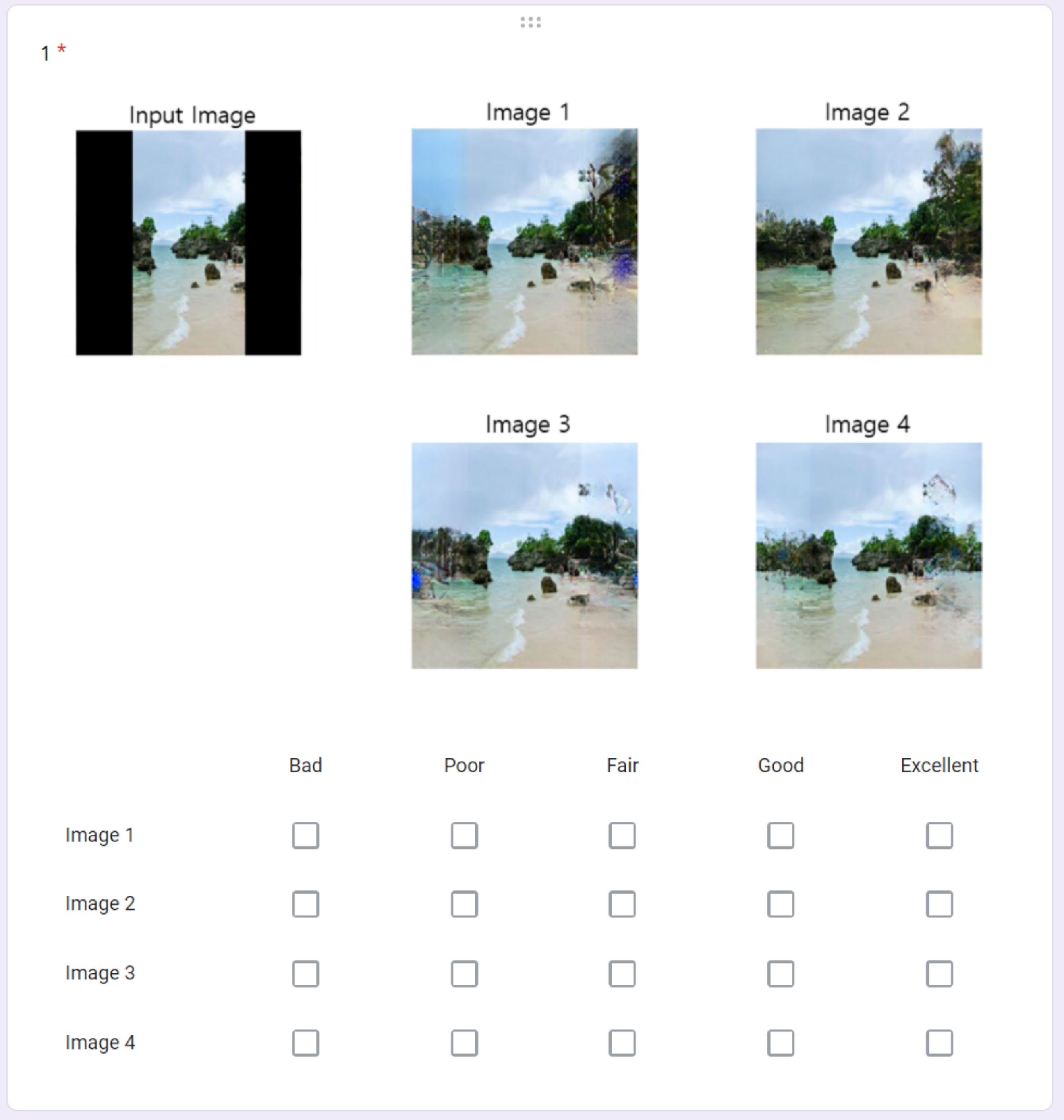


Figure 11. Qualitative results for conventional and proposed methods on Beach dataset.



1 *

	Bad	Poor	Fair	Good	Excellent
Image 1	□	□	□	□	□
Image 2	□	□	□	□	□
Image 3	□	□	□	□	□
Image 4	□	□	□	□	□

Figure 12. An example of MOS Survey Form we used.