Supplementary Material: Outpainting by Queries

A Additional Quantitative Results

	Methods	Scenery			Building Facades			WikiArt		
		FID↓	IS↑	PSNR↑	FID↓	IS↑	PSNR↑	FID↓	IS↑	PSNR↑
1×	Lower Bound	160.174	3.595	9.569	123.678	4.356	9.810	139.956	5.073	10.215
	SRN	45.296	3.540	22.433	34.058	4.722	18.839	65.675	4.933	**20.467**
	NSIPO	35.606	3.475	21.630	33.140	4.529	18.460	30.338	6.231	18.929
	IOH	23.410	3.578	22.839	33.525	4.739	18.812	24.539	6.679	19.808
	Uformer	23.216	3.691	23.054	32.228	4.651	18.892	18.808	7.466	19.708
	QueryOTR	**20.366**	**3.955**	**23.604**	**22.378**	**4.978**	**19.680**	**14.955**	**7.896**	20.388
2×	Lower Bound	201.871	2.097	7.868	196.650	2.875	8.047	230.893	2.477	8.557
	SRN	97.989	2.724	18.459	75.121	3.837	15.431	139.395	3.045	16.759
	NSIPO	69.683	3.235	17.701	65.319	3.771	15.287	67.880	4.888	15.721
	IOH	45.108	3.047	18.846	72.053	3.727	15.519	66.953	5.065	16.127
	Uformer	50.605	3.099	18.934	71.306	3.924	15.626	51.263	6.098	15.936
	QueryOTR	**39.237**	**3.431**	**19.358**	**41.273**	**4.547**	**16.213**	**43.757**	**6.341**	**17.074**
3×	Lower Bound	227.268	1.991	7.242	223.224	2.378	7.384	260.623	2.258	7.919
	SRN	141.040	2.483	16.141	114.016	3.312	13.777	181.394	2.407	14.620
	NSIPO	101.411	**3.131**	15.384	92.041	3.628	13.741	94.176	4.325	14.159
	IOH	67.591	2.723	16.351	104.337	2.956	13.913	104.032	4.190	13.943
	Uformer	76.318	2.799	16.374	105.539	3.315	14.065	79.322	**5.954**	13.411
	QueryOTR	**60.977**	3.114	**16.864**	**64.926**	**4.612**	**14.316**	**69.951**	5.683	**15.294**
	Up Bound	0	4.184	$+\infty$	0	5.660	$+\infty$	0	8.779	$+\infty$

Table S1: Quantitative results follow our setting that **replacing the center region with input sub-images surrounded by the extrapolated parts**.

The comparative methods are all based on image-to-image translation, which need to reconstruct the input sub-image, whilst the sequence-to-sequence based method QueryOTR does not need to reconstruct the input sub-image only outputting the extrapolated regions. In the main manuscript, we report the best results of comparative methods by keeping the reconstructed regions, which were consistent with their original settings. All things being equal, Table S1 reports the results following our setting that replaces the center region with the input sub-image. We additionally report the lower bound of each metric by filling the extrapolated regions with zero pixel values. As shown in Table S1, our proposed QueryOTR outperforms other methods in most cases, indicating that the higher performance of our method contributes little on the use of the input sub-image. Instead, the results demonstrate the superiority of our method on generating the

extrapolated regions. On the other hand, all the comparative methods have an improvement of IS and PSNR metrics due to replacing the center regions with the input sub-image.

B Details of Datasets

Scenery is a natural scenery dataset consisting of about 5,000 images in the training set and 1,000 images in the testing set. The images are very diverse and complicated, which contains natural scenes, e.g., snow, valley, seaside, riverbank, sky, and mountain.
Building Facades is a city scenes dataset consisting of about 16,000 and 1,500 images for training set and testing set respectively. It contains building architecture and city scenes.
WikiArt is a fine-art paintings dataset obtained from the wikiart.org website. We use the split manner of genres datasets, which contains 45,503 training images and 19,492 testing images.

C Inference Time

The comparison of inference time can be referred in Table S2. Due to the simple but effective design of QueryOTR, our framework is almost three times faster than Uformer which is also engaged with a vision transformer (Swin Transformer) architecture.

Method	SRN	NSIPO	IOH	Uformer	QueryOTR
Time usage (ms/image)	11.960	44.190	4.160	46.810	13.345

Table S2: Comparison on inference time.

D Hard Examples

We illustrate some hard examples that QueryOTR can work significantly better than the other methods. As shown in Figure S1, extrapolating the images with simple colour stripes is very challenging, which requires the network to recognize the pattern and mimic it, especially when such samples are not enough in the training set. The CNN-based methods have limitations to generalize well on such samples, whilst the transformer-based Uformer can generate colorful lines but not straight. In contrast, QueryOTR takes advantage of querying the input sub-image to generate color and straight lines, generating much better images.

E More Qualitative Results

We present more comparative results for one-step and multi-step outpainting. In Figure S2, we show additional results on Scenery and WikiArt datasets. Similarly, in Figure S3, we provide more results on Building Facades dataset compared with other methods. Meanwhile, we visualize the results conducted by QueryOTR in Figure S4.

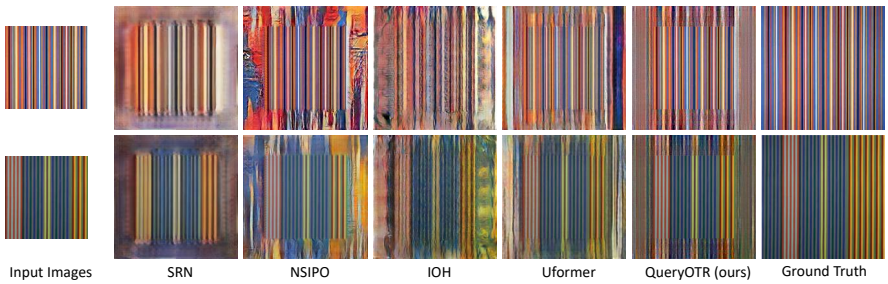

Fig. S1: Visualization of some hard examples in the test set of WikiArt dataset.

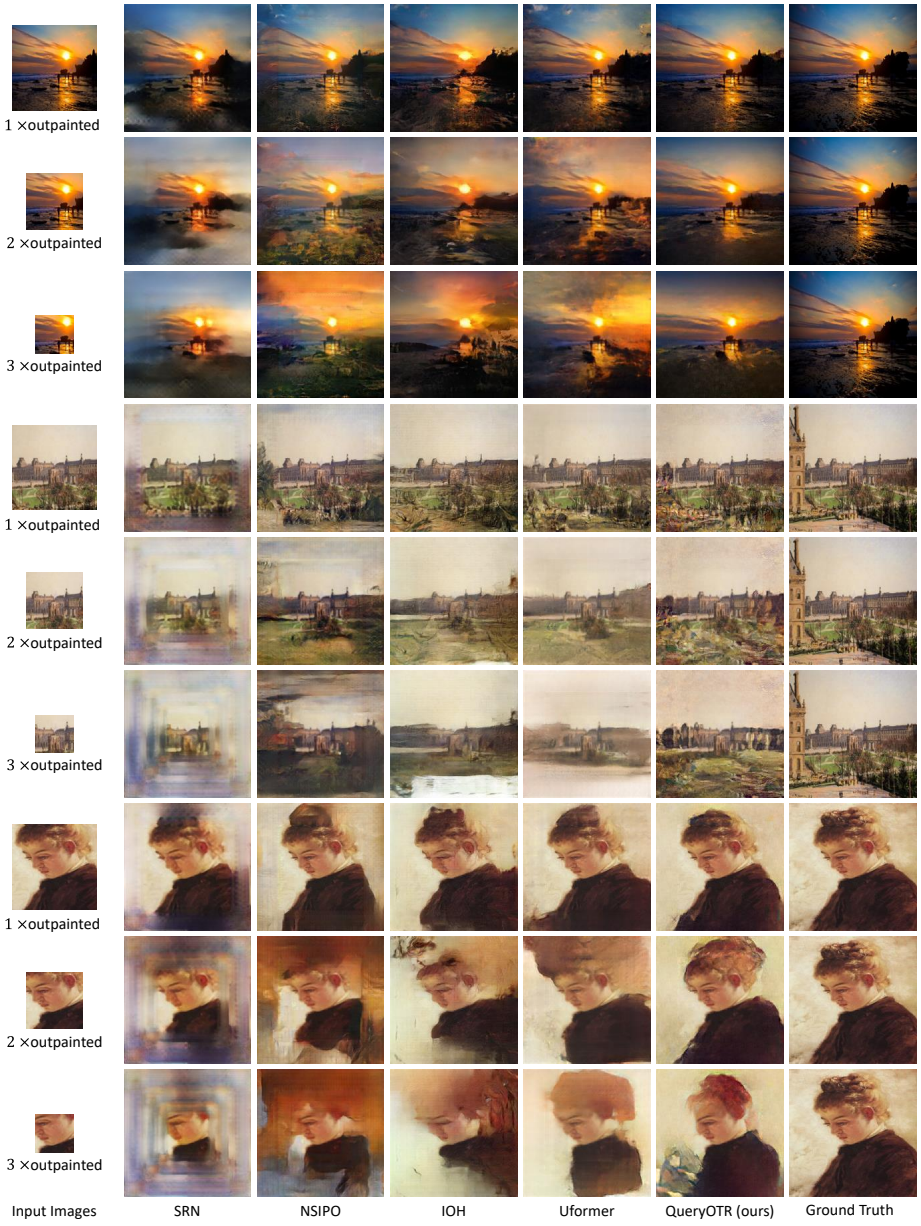

Fig. S2: Qualitative comparison results on Scenery and WikiArt datasets.

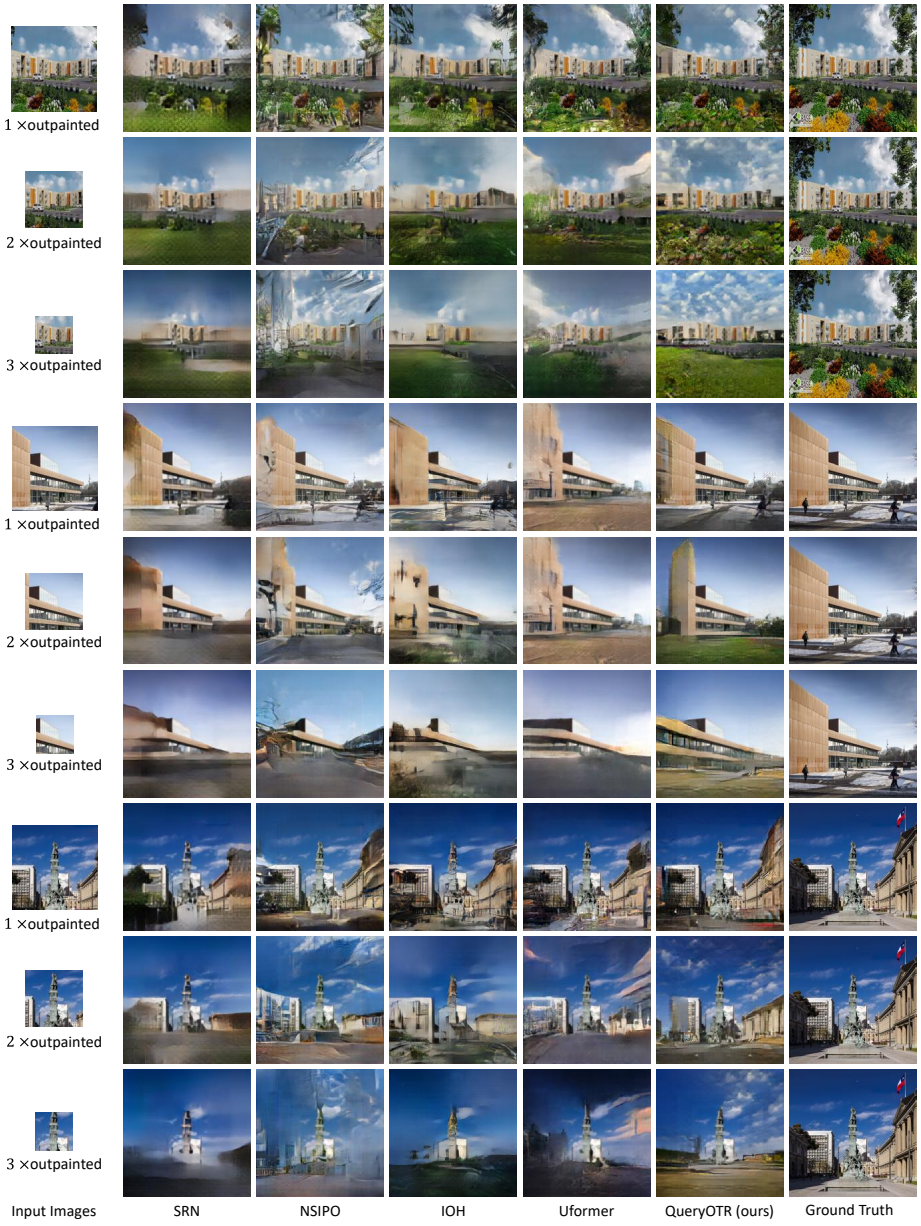

Fig. S3: Qualitative comparison results on Building Facades dataset.

Fig. S4: Visualization of QueryOTR one-step outpainting.