EUROPEAN CONFERENCE ON COMPUTER VISION
TEL AVIV 2022
October 23-27, 2022, Tel Aviv

Outpainting by Queries

Kai Yao [1,2], Penglei Gao [1,2], Xi Yang [1],
Jie Sun [1], Rui Zhang [1], Kaizhu Huang [3]
[1] Xi'an Jiaotong-Liverpool university
[2] University of Liverpool [3] Duke Kunshan University

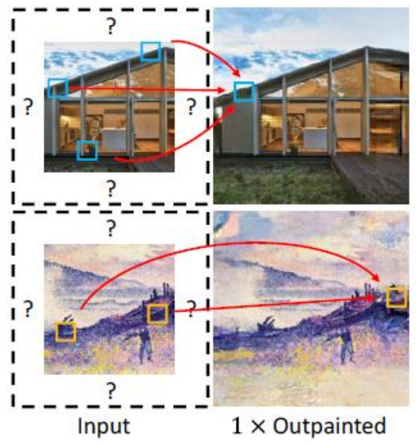
Introduction

The challenges of image outpainting:

(a) Determining where the missing features should be located relative to the output's spatial locations for both nearby and faraway features.

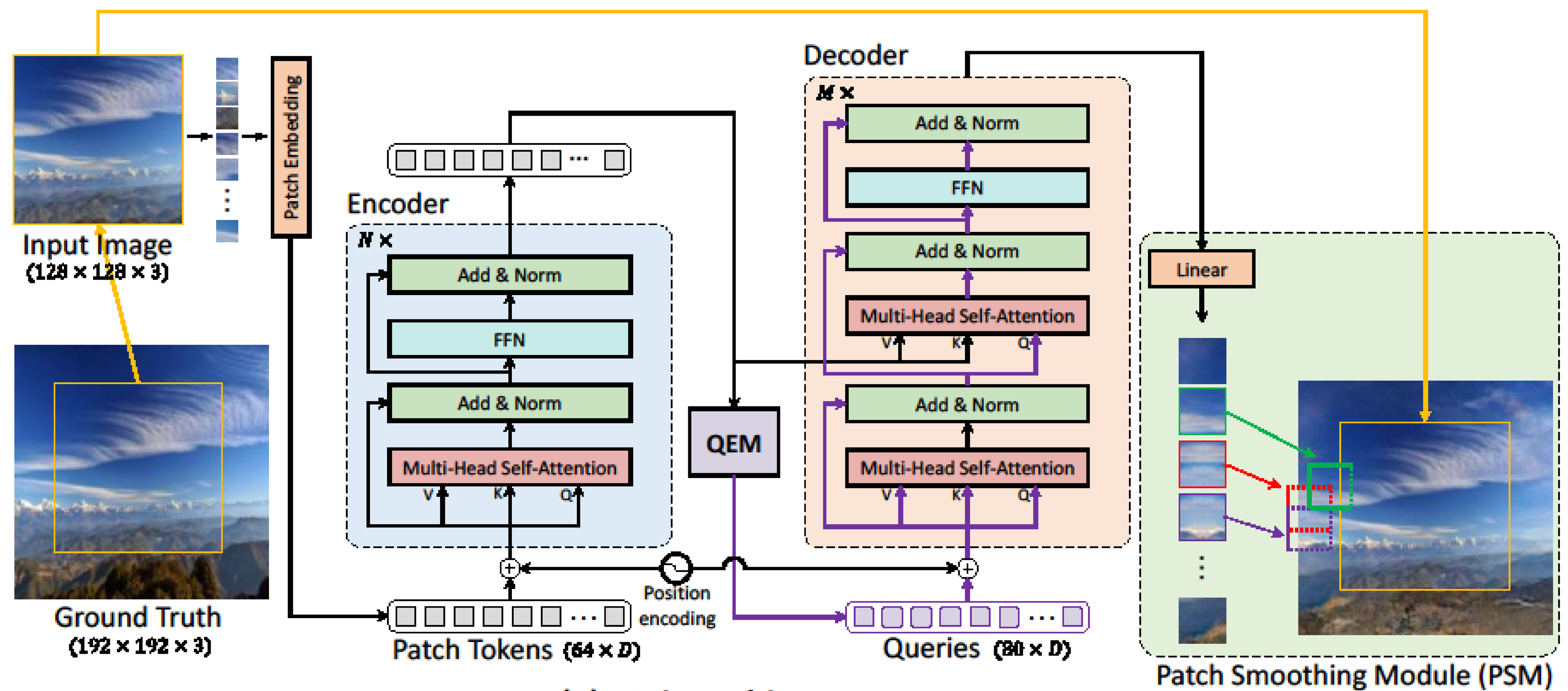
(b) Guaranteeing that the extrapolated image has a realistic appearance with reasonable content and a consistent structural layout with the conditional sub-image.

(c) The borders between extrapolated regions and the original sub-image should be smooth and seamless.

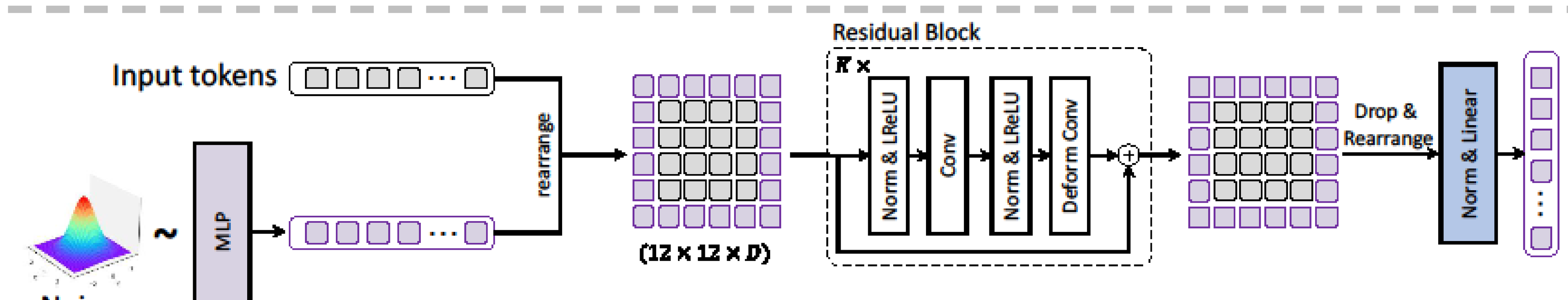


We reconsider the outpainting problem as a patch-wise sequence-to-sequence autoregression problem. We propose Query Expansion Module and Patch Smoothing Module to solve the slow convergence problem in pure transformers and to generate realistic extrapolated images smoothly and seamlessly.

Architecture



(a) Main architecture



(b) Query Expansion Module (QEM)

Given an image $x \in \mathbb{R}^{H \times W \times 3}$, we aim to extrapolate outside contents beyond the image boundary with extra M-pixels. The generator will produce an image $\hat{x} \in \mathbb{R}^{(H+2M) \times (W+2M) \times 3}$. The goal is to predict the extra sequence $\{x_p^{L+1}, x_p^{L+2}, \ldots, x_p^{L+R}\}$, where $x_p^i \in \mathbb{R}^{p^2 \cdot 3}$.
The proposed QEM is designed to speed up the convergence of pure transformer by generating the expanded queries for the transformer decoder. We predict the decoders' queries conditioned on encoders' features, and take advantage of CNN's inductive bias to accelerate the convergence.
PSM is designed to mitigate the artifacts issue by considering the neighboring patches' content enabling the output sequence to have same length but less effect as the predefined grids.

Experimental Results

	Methods	Scenery			Building Facades			WikiArt		
		FID↓	IS↑	PSNR↑	FID↓	IS↑	PSNR↑	FID↓	IS↑	PSNR↑
1×	SRN	47.781	2.981	22.440	38.644	3.862	18.588	76.749	3.629	20.072
	NSIPO	25.977	3.059	21.089	30.465	4.153	18.314	22.242	5.600	18.592
	IOH	32.107	2.886	22.286	49.481	3.924	18.431	40.184	4.835	19.403
	Uformer	20.575	3.249	23.007	30.542	4.189	18.828	15.904	6.567	19.610
	QueryOTR	**20.366**	**3.955**	**23.604**	**22.378**	**4.978**	**19.680**	**14.955**	**7.896**	**20.388**
2×	SRN	83.772	2.349	18.403	74.304	3.651	15.355	137.997	3.039	16.646
	NSIPO	45.989	2.606	17.733	58.341	3.669	15.262	51.668	4.591	15.679
	IOH	44.742	2.655	18.739	76.476	3.456	15.443	75.070	4.289	16.056
	Uformer	39.801	2.920	18.920	63.915	3.798	15.612	**41.107**	5.900	15.947
	QueryOTR	**39.237**	**3.431**	**19.358**	**41.273**	**4.547**	**16.213**	43.757	**6.341**	**17.074**
3×	SRN	115.193	2.087	16.123	110.036	2.938	13.693	181.533	2.504	14.609
	NSIPO	64.457	2.405	15.606	81.301	3.431	13.791	75.785	4.225	14.257
	IOH	**58.629**	2.432	16.307	95.068	2.790	13.894	108.328	3.728	13.919
	Uformer	60.497	2.638	16.379	93.888	3.388	14.051	72.923	**5.904**	13.464
	QueryOTR	60.977	**3.114**	**16.864**	**64.926**	**4.612**	**14.316**	**69.951**	5.683	**15.294**

Table 1: Quantitative results of one-step and multi-step outpainting. Best and second best results are **boldface** and underlined. 1× represents one step outpainting, while 2× and 3× denote two- and three-step outpainting respectively.

Pretrained *Enc.*	M	FID↓	IS↑
-	4	22.784	3.751
✓	2	20.731	3.931
✓	4	**20.366**	**3.955**
✓	8	20.373	3.852

(a) Ablation of the pretrained ViT-base encoder and the number of transformer decoder layers M.

	FID↓	IS↑
w/o $\mathcal{L}_{rec}$ & $\mathcal{L}_{perceptual}$	38.009	3.433
w/o $\mathcal{L}_{rec}$	31.282	3.744
w/o $\mathcal{L}_{perceptual}$	33.380	3.510
QueryOTR (baseline)	**20.366**	**3.955**

(b) Impact of $\mathcal{L}_{rec}$ and $\mathcal{L}_{perceptual}$ contribute to the overall performance. The model is default trained with three losses.

	FID↓	IS↑
w/o QEM	36.967	3.642
QEM w/o Noise	23.444	3.728
QEM w/o DC [48]	23.530	**3.775**
w QEM	**22.784**	3.751

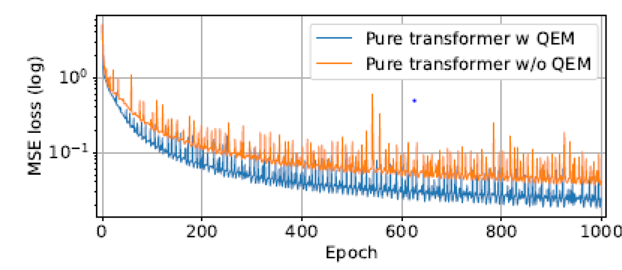
(c) Impact of proposed Query Expansion Module (QEM) and its key internal components.

PSM	Per-Patch *Norm.*	FID↓	IS↑
-	-	51.945	3.801
-	✓	31.073	3.753
✓	-	22.501	3.707
✓	✓	**20.366**	**3.955**

(d) Effect of the proposed Patch Smoothing Module (PSM) and per-patch image normalization.

Table 2: Ablation studies validated on Scenery dataset.

Our QueryOTR could generate more realistic images with vivid details and enrich the contents of the generated regions marked in white box. Furthermore, our method could weaken the sense of edges between the generated regions and input sub-image.



(a) Autoregression w/ and w/o QEM

QEM significantly speeds up the convergence (about 3.3 times faster than that without QEM.

(a) w/o QEM (b) w/o PSM

Conclusion

We proposed a novel hybrid query-based encoder-decoder transformer framework to extrapolate visual context all-side around a given image. The QEM helps to accelerate the transformer model convergence and PSM contributes to generate seamless extrapolated images realistically and smoothly.

Acknowledgments:The work was partially supported by the following: National Natural Science Foundation of China under no.61876155; Jiangsu Science and Technology Programme under no.BE2020006-4; Key Program Special Fund in XJTLU under no.KSF-T-06 and no.KSF-E-37; Research Development Fund in XJTLU under no.RDF-19-01-21.