Taming Transformers for High-Resolution Image Synthesis

Patrick Esser* Robin Rombach* Björn Ommer

{firstname.lastname}@iwr.uni-heidelberg.de

HCI, IWR, Heidelberg University

*equal contribution

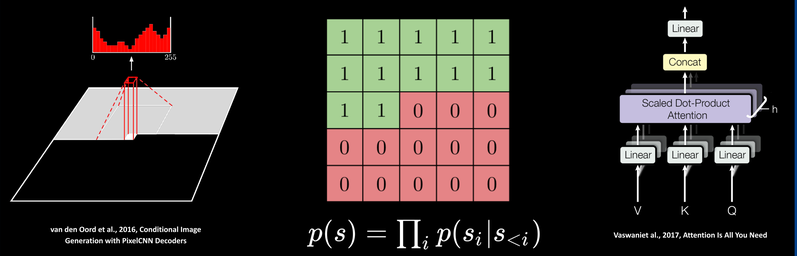
CVPR VIRTUAL JUNE 19-25

Generative Models: Applications

Guided Synthesis | Image Editing | One-Shot Learning | Generative Pretraining

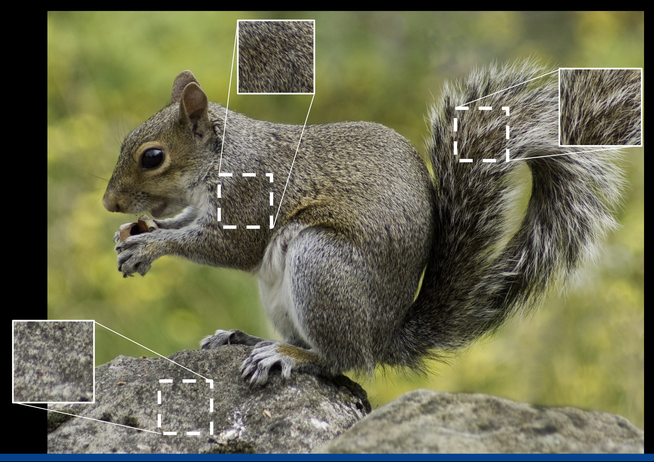
Autoregressive Models: Complexity

Local Convolution + Autoregressive Mask + Global Attention

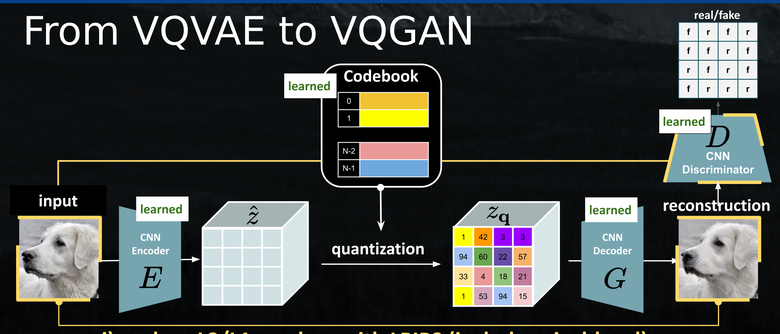


$\mathcal{O}$(#pixels · kernel size) $\mathcal{O}$(#pixels²)

Textures: Imperceptible Details



From VQVAE to VQGAN

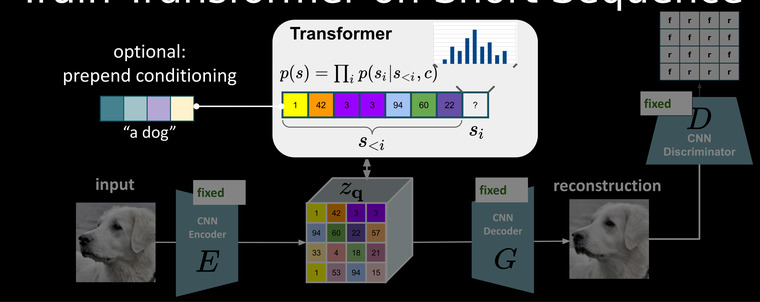


i) replace L2/L1 rec. loss with LPIPS (includes pixel-level)

ii) add (patch-wise) Discriminator to favor realism over perfect reconstruction

Perceptual Compression

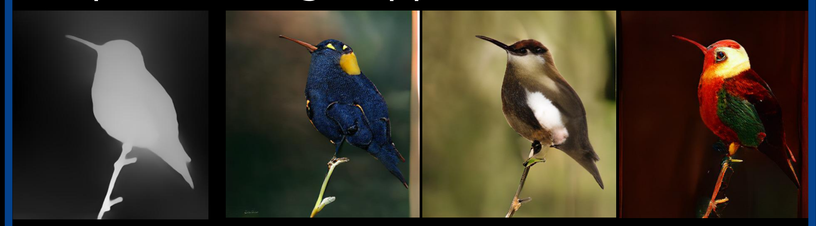
Train Transformer on Short Sequence



Synthesis of Megapixel Landscapes

Class-Conditional ImageNet

Depth-to-Image Applications



Code: https://compvis.github.io/taming-transformers/

- ✓ do not waste capacity on imperceptible details of images
- ✓ reduce computational cost by a factor 16
- ✓ synthesize high-resolution images with global attention
- ✓ competitive with state of the art