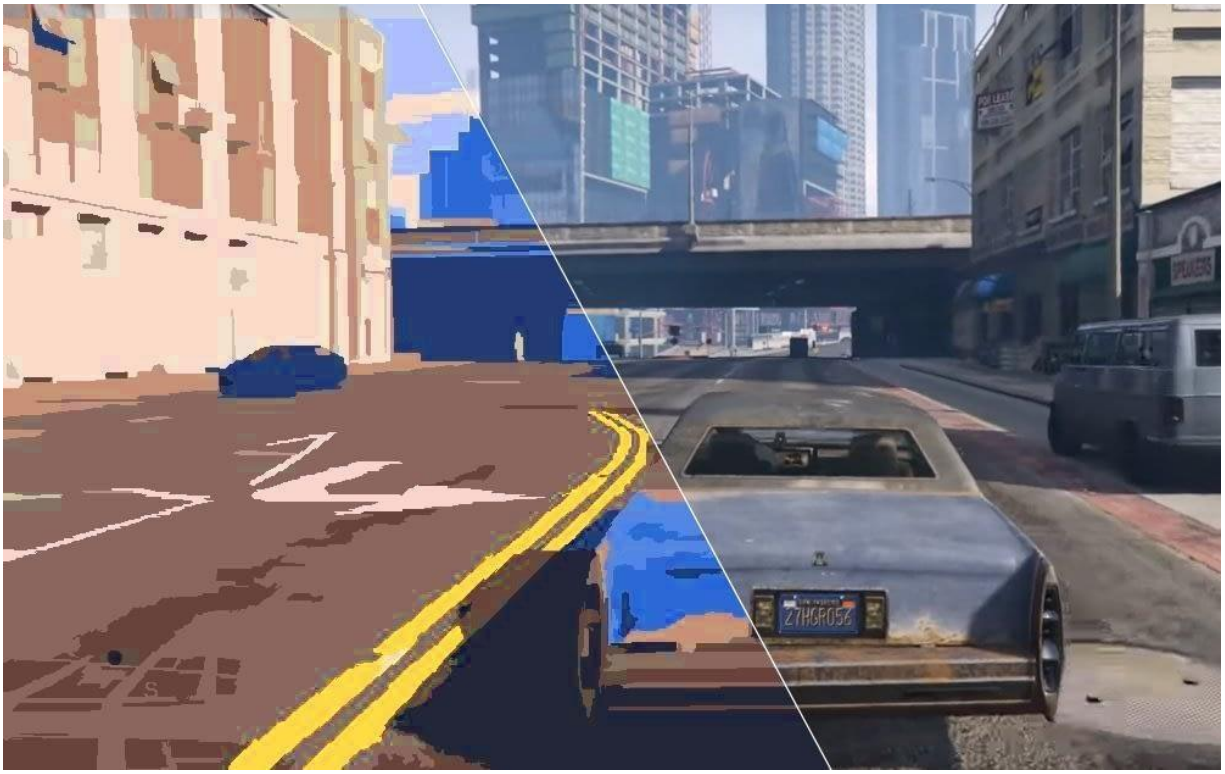
Neural Rendering: How Low Can You Go In Terms of Input?

By Martin Anderson



First published ***May 13th, 2021*** *at:*

https://www.unite.ai/neural-rendering-low-resolution-input-intel/ | *Web-archived version*

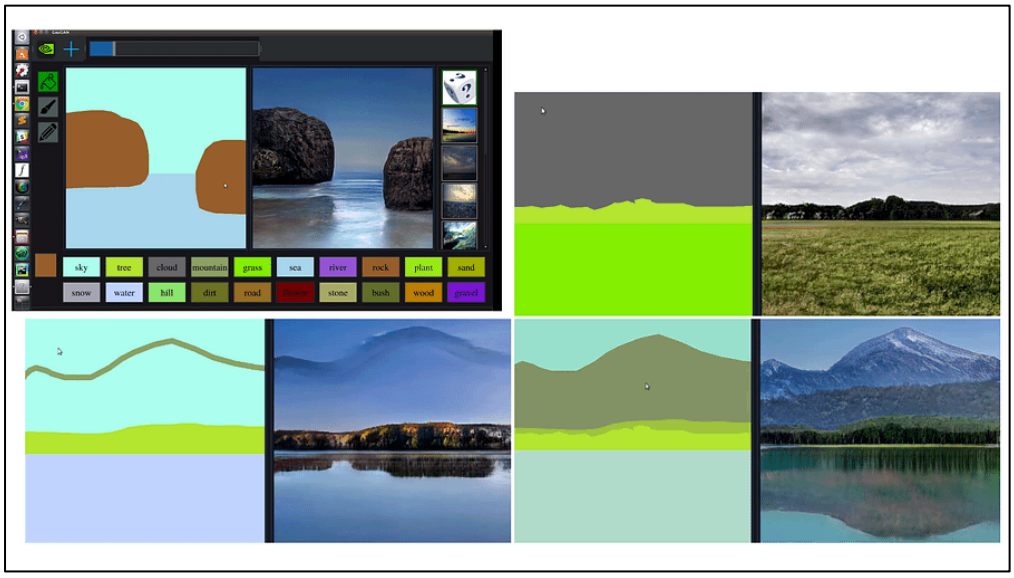
Yesterday some extraordinary new work in neural image synthesis caught the attention and the imagination of the internet, as Intel researchers revealed a new method for enhancing the realism of synthetic images.

The system, as demonstrated in a video from Intel, intervenes directly into the image pipeline for the Grand Theft Auto V video game, and automatically enhances the images through an image synthesis algorithm trained on a convolutional neural network (CNN), using real world imagery from the Mapillary dataset, and swapping out the less realistic lighting and texturing of the GTA game engine.

Commenters, in a wide range of reactions in communities such as Reddit and Hacker News, are positing not only that neural rendering of this type could effectively replace the less photorealistic output of traditional games engines and VFX-level CGI, but that this process could be achieved with far more basic input than was demonstrated in the Intel GTA5 demo — effectively creating 'puppet' proxy inputs with massively realistic outputs.

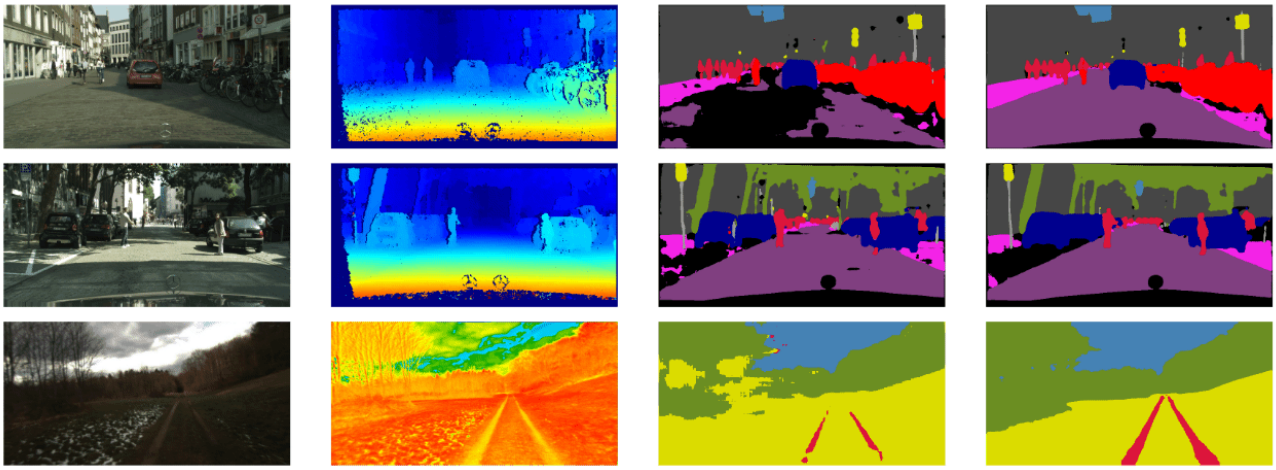
Paired Datasets

The principle has been exemplified by a new generation of GAN and encoder/decoder systems over the last three years, such as NVIDIA's GauGAN, which generates photorealistic scenic imagery from crude daubs.



Effectively this principle flips the conventional use of semantic segmentation in computer vision from a passive method that allows machine systems to identify and isolate observed objects into a creative input,

where the user 'paints' a faux semantic segmentation map and the system generates imagery that's consistent with the relationships it understands from having already classified and segmented a particular domain, such as scenery.



A machine learning framework applies semantic segmentation to various exterior scenes, providing the architectural paradigm that permits the development of interactive systems, where the user paints a semantic segmentation block and the system infills the block with apposite imagery from a domain-specific dataset, such as Germany's Mapillary street view set, used in Intel's GTA5 neural rendering demo. Source: http://ais.informatik.uni-freiburg.de/publications/papers/valada17icra.pdf

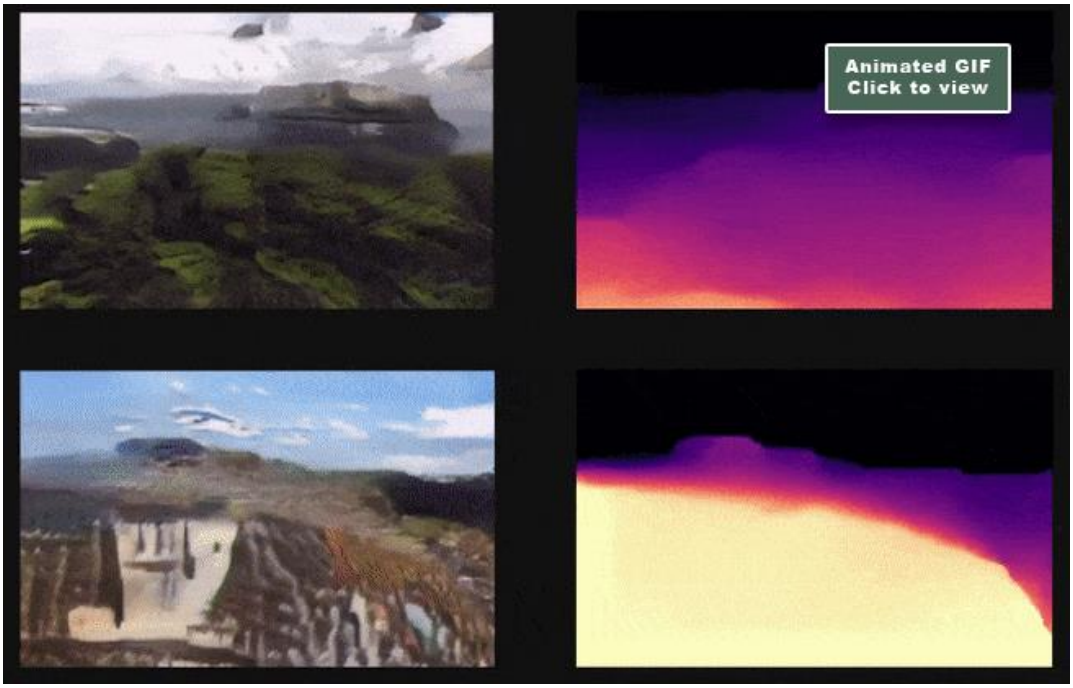
Paired dataset image synthesis systems work by correlating semantic labels on two datasets: a rich and full-fledged image set, either generated from real-world imagery (as with the Mapillary set used to enhance GTA5 in yesterday's Intel demo) or from synthetic images, such as CGI images.

Paired dataset examples for an image synthesis system designed to create neural-rendered characters from clumsy sketches. On the left, samples from the CGI dataset. Middle, corresponding samples from the 'sketch' dataset. Right, neural renders that have translated sketches back into high-quality images. Source: https://www.youtube.com/watch?v=miLIwQ7yPkA

Exterior environments are relatively unchallenging when creating paired dataset transformations of this kind, because protrusions are usually quite limited, the topography has a limited range of variance that can be comprehensively captured in a dataset, and we don't have to deal with creating artificial people, or negotiating the Uncanny Valley (yet).

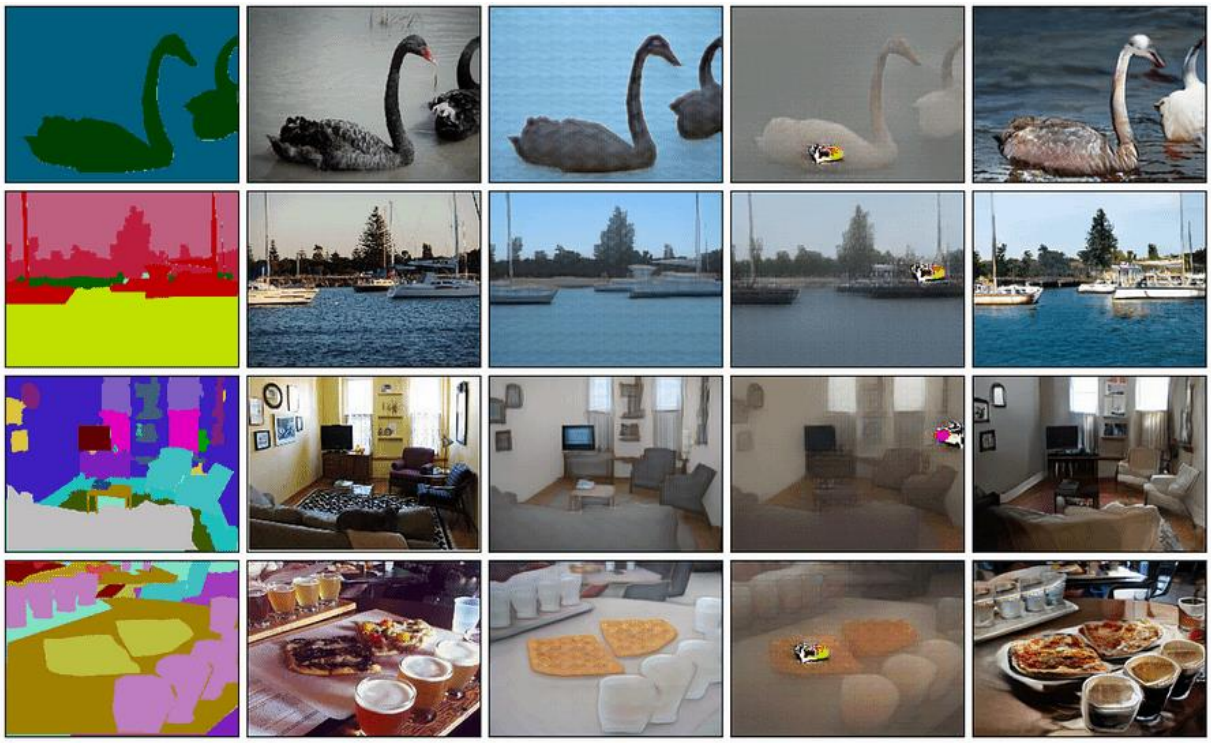
Inverting Segmentation Maps

Google has developed an animated version of the GauGAN schema, called Infinite Nature, capable of deliberately 'hallucinating' continuous and never-ending fictitious landscapes by translating fake semantic maps into photorealistic imagery via NVIDIA's SPADE infill system:



Source: https://www.youtube.com/watch?v=oXUf6anNAtc

However, Infinite Nature uses a single image as a starting point and uses SPADE merely to paint in the missing sections in successive frames, whereas SPADE itself creates image transforms directly from segmentation maps.



Source: https://nvlabs.github.io/SPADE/

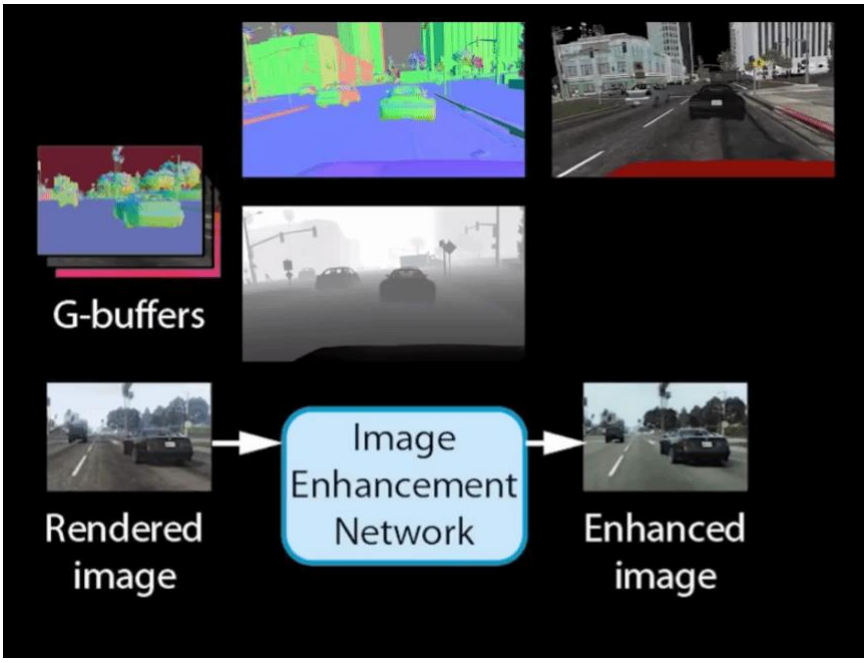
It is this capacity that seems to have stirred admirers of the Intel Image Enhancement system – the possibility of deriving very high quality photorealistic imagery, even in real time (eventually), from extremely crude input.

Replacing Textures and Lighting With Neural Rendering

In the case of the GTA5 input, some have wondered whether any of the computationally expensive procedural and bitmap texturing and lighting from the game engine output is really going to be necessary in future neural rendering systems, or whether it might be possible to transform low-resolution, wireframe-level input into photorealistic video that outperforms the shading, texturing and lighting capabilities of game engines, creating hyper-realistic scenes from 'placeholder' proxy input.

It might seem obvious that game-generated facets such as reflections, textures, and other types of environmental detail are essential sources of information for a neural rendering system of the type demonstrated by Intel. Yet it has been some years since NVIDIA's UNIT (UNsupervised Image-to-image Translation Networks) demonstrated that only the domain is important, and that even sweeping aspects such as 'night or day' are essentially issues to be handled by style transfer:

In terms of required input, this potentially leaves the game engine only needing to generate base geometry and physics simulations, since the neural rendering engine can over-paint all other aspects by synthesizing the desired imagery from the captured dataset, using semantic maps as an interpretation layer.



Intel's system enhances a completely finished and rendered frame from GTA5, adding segmentation and evaluated depth maps — two facets which could potentially be supplied directly by a stripped-down game engine. Source: https://www.youtube.com/watch?v=P1IcaBn3ej0

Intel's neural rendering approach involves the analysis of completely rendered frames from the GTA5 buffers, and the neural system has the added burden of creating both the depth maps and the segmentation maps. Since depth maps are implicitly available in traditional 3D pipelines (and are less demanding to generate than texturing, ray-tracing or global illumination), it might be a better use of resources to let the game engine handle them.

Stripped-Down Input for a Neural Rendering Engine

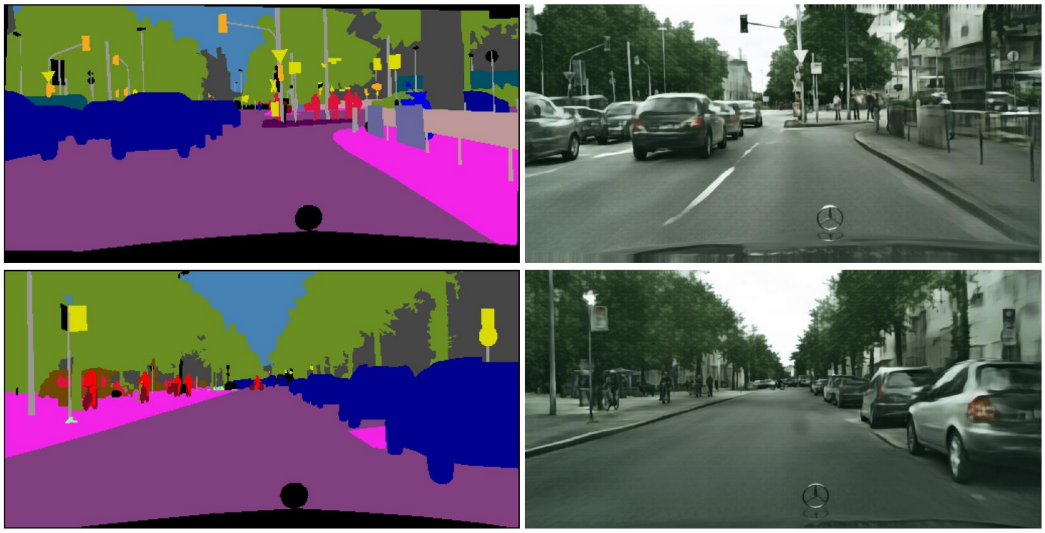
The current implementation of the Intel image enhancement network, therefore, may involve a great deal of redundant computing cycles, as the game engine generates computationally expensive texturing and lighting which the neural rendering engine does not really need. The system seems to have been designed in this way not because this is necessarily an optimal approach, but because it is easier to adapt a neural rendering engine to an existing pipeline than to create a new game engine that is optimized to a neural rendering approach.

The most economical use of resources in a gaming system of this nature could be complete co-opting of the GPU by the neural rendering system, with the stripped-down proxy input handled by the CPU.

Furthermore, the game engine could easily produce representative segmentation maps itself, by turning off all shading and lighting in its output. Additionally, it could supply video at a far lower resolution than is normally required of it, since the video would only need to be broadly representative of the content, with high resolution detail being handled by the neural engine, further freeing up local compute resources.

Intel ISL's Prior Work With Segmentation>Image

The direct translation of segmentation to photorealistic video is far from hypothetical. In 2017 Intel ISL, creators of yesterday's furor, released initial research capable of performing urban video synthesis directly from semantic segmentation.

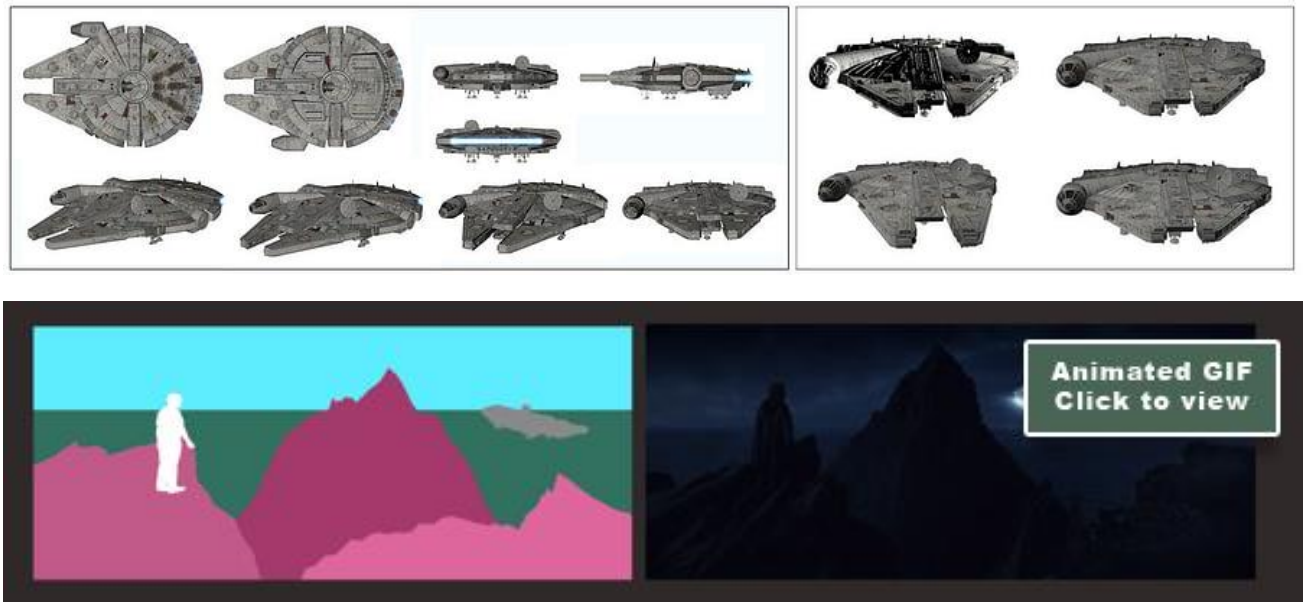


Intel ISL's segmentation to image work from 2017. Source: https://awesomeopensource.com/project/CQFIO/PhotographicImageSynthesis

In effect, that original 2017 pipeline has merely been extended to fit GTA5's fully rendered output.

Neural Rendering in VFX

Neural rendering from artificial segmentation maps also seems to be a promising technology for VFX, with the possibility of directly translating very basic videograms directly into finished visual effects footage, by generating domain-specific datasets taken either from models or synthetic (CGI) imagery.



A hypothetical neural rendering system, where extensive coverage of each target object is abstracted into a contributing dataset, and where artificially-generated segmentation maps are used as the basis for full-resolution photorealistic output. Source: https://rossdawson.com/futurist/implications-of-ai/comprehensive-guide-ai-artificial-intelligence-visual-effects-vfx/

The development and adoption of such systems would shift the locus of artistic effort from an interpretive to a representative workflow, and elevate domain-driven data gathering from a supporting to a central role in the visual arts.


Enhancing Photorealism Enhancement
GTA V
Ours