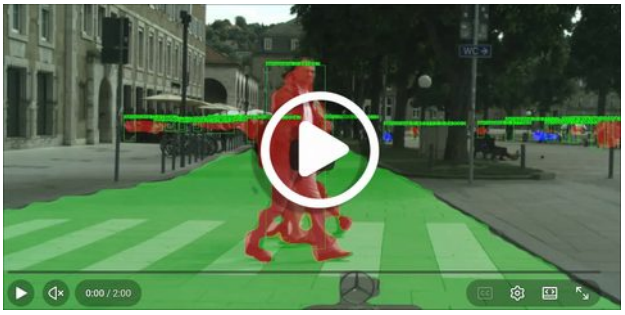
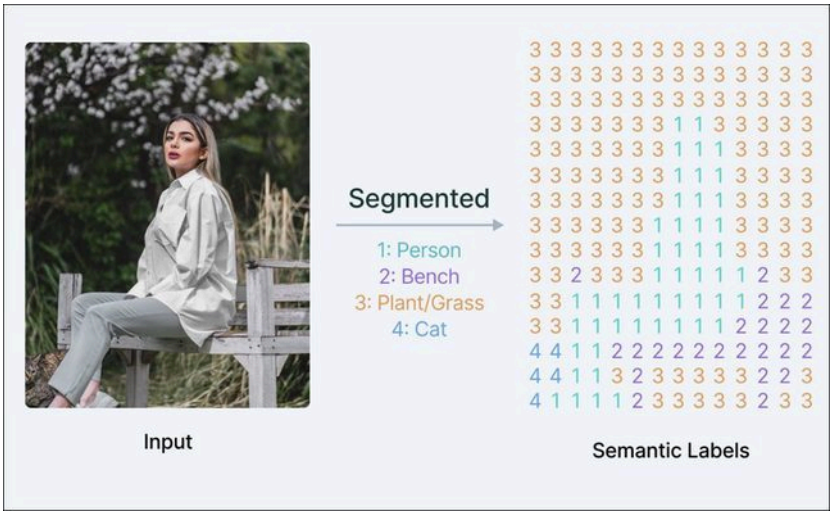
Semantic Segmentation

Originally published February 1, 2023 at https://blog.metaphysic.ai/semantic-segmentation/

Semantic segmentation is the process of individuating objects (people, cars, faces, or anything else the network has been trained to recognize) inside an image or video frame, and delineating the borders of those objects, so that it's clear where the boundaries are.



In practical terms, the areas of the pixel grid-map that are occupied by such 'semantically meaningful' material are assigned related values, which at the very least will be a 0 or a 1 (i.e., 'cat' | 'not cat'). A system designed to recognize more than one type of item will have diverse values for each object recognized:



Different facets of an image get their own number in this hypothetical multi-subject semantic segmentation application. Source: https://www.v7labs.com/blog/semantic-segmentation-guide

In a mono-subject semantic segmentation application, the system is only looking for a '1' value, and is indiscriminate about how many examples of the target subject there are in the image. For instance, an infra-red system designed to recognize foxes will assign '1' to pixels that represent any number of onscreen foxes, and all you can determine from such a system is that there is at least one fox in the shot.

Source: https://mdpi-res.com/d_attachment/animals/animals-11-01723/article_deploy/animals-11-01723.pdf

On the other hand, an *instance segmentation* system is designed to individuate multiple examples of the same type of recognized object, and will be able to categorize 'fox_1', 'fox_2', and so on.

CLICK TO VIEW ANIMATED GIF

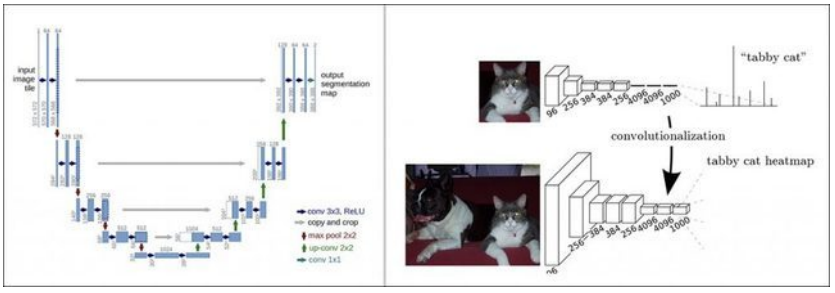
A semantic segmentation framework that incorporates instance recognition, such as You-Only-Look-Once (YOLO), pictured here, can individuate multiple instances of a class or label. Source: https://www.youtube.com/watch?v=tq0GI4FahWU

As we can see in the fox and samurai examples above, the most basic delineation of semantic segmentation is to define the outermost margins of the recognized object, providing a *bounding box*. Since the entire pixel grid has to be traversed in order to define 'not zero' pixels (i.e., a recognized subject), creating a more complex defining outline, as in the lower image, is not a great leap, except that it may require greater processing power to render in real time. Typically, there is a trade-off between efficiency and accuracy, usually expressed in the lowering of the frame rate (we can see that the lower and more complex semantic segmentation example of the two images above has been captured at 9fps).

CNNs in Semantic Segmentation

A typical semantic segmentation framework will at the very least have made use of a Convolutional Neural Network (CNN), which is capable of learning complex features from images, including broad shapes, delineations and textures.

However, a basic CNN only produces a single output for each image, whereas a Fully Convolutional Neural Network (FCNN), introduced specifically for the semantic segmentation task, by researchers at UC Berkeley in 2015, can accommodate arbitrary inputs and outputs, enabling heatmap-style visualization, wherein individual facets of an image can be isolated, and multiple identifications counted, even across classes and labels.



The U-Net network, the naming of which is clear from the illustration on the left, enables a more complex throughput of variable sizes of data in a CNN. Source: https://arxiv.org/pdf/1411.4038.pdf

Technically, an FCNN is actually a 'down-sized' version of a CNN, since it lacks fully connected layers, and is specifically designed to process subsampling and upsampling operations. However, this optimized approach makes it a powerful architecture for semantic segmentation.

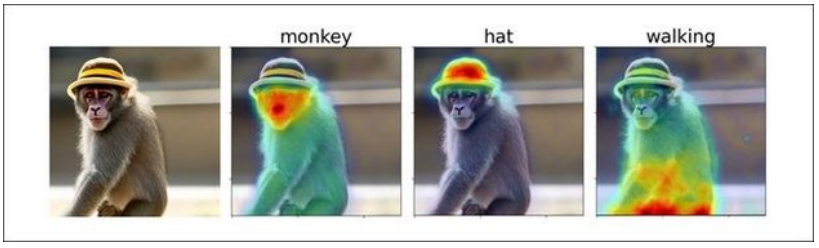
Semantic Segmentation in Image Synthesis

In the emerging age of multimodal image synthesis systems such as latent diffusion, which is powered by the connection between class labels and pixel data, there is growing interest in using semantic segmentation as a means of helping generative systems to isolate facets of an image. One application for this functionality is to help to disentangle a class or labeled content from the context in which it is sitting in the image, so that systems which are training on labeled data do not engage in shortcut learning (i.e., do not over-associate dogs with grass and pavements, or beachwear with beaches, etc.), but only learn the actual content of the class, both as a lexical term and as an isolated group of pixels.

Another way in which semantic segmentation could aid the AI-enabled image editing systems of the future is by recognizing and isolating subjects inside an image so that transformations applied to the subject do not excessively change their context or environment.

On the surface, this sounds like a souped-up version of Photoshop's lasso tool from the early 1990s, except that the isolation is intended to take place in the latent space of the generative system, so that all transformations will have occurred by the time the content is visible to the viewer. In this way, the synergistic effects of a requested transformation can be considered in the more ductile features embedded in a neural environment, and not in the rigid pixel space of an explicitly-rendered image.

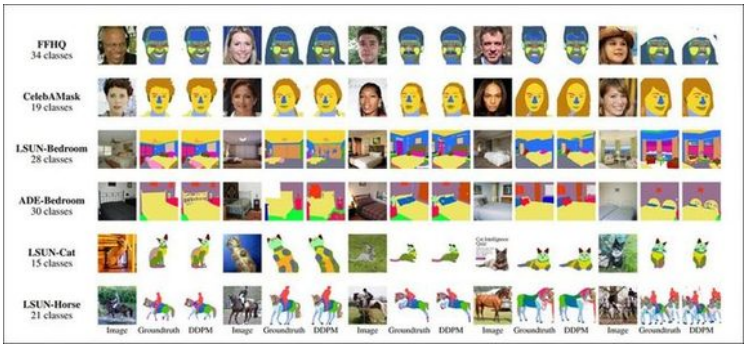
One academic/industry collaboration from late 2022 has adopted semantic segmentation principles to create a kind of Gradient Class Activation Maps (Grad-CAM, a tool typically used in Generative Adversarial Networks) system for Stable Diffusion, wherein heat-maps indicate to the viewer which parts of a text-prompt influenced various segments of a generated image.



DAMM heat-maps from the COCO dataset could signal a new branch of semantic segmentation, applicable to the strictures of latent diffusion architectures. Source: https://arxiv.org/pdf/2210.04885.pdf

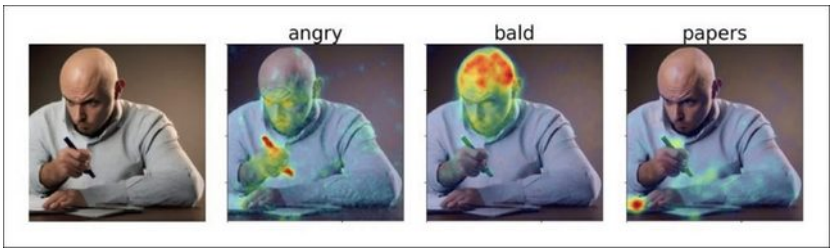
In such a case, the language component of the text-prompt, and the way it activates features trained on pertinent and corresponding text (i.e., trained labels associated with apposite imagery) is being used as a kind of 'Barium meal' to highlight and delineate content, instead of training an external system to recognize pixel groupings with familiar patterns that correspond to a class.

However, this particular new and emerging thread of research seems unlikely to take hold in the real-time semantic segmentation sector, currently dominated by the lightweight You-Only-Look-Once (YOLO) series, since inference time seems likely to stay quite high even for purely off-line neural queries, with scant hope of 'injecting' novel live information into a diffusion system and obtaining a usable response time.
Nonetheless, this is an active line of research; one 2022 project from Yandex proposed a diffusion-based semantic segmentation model.



A diffusion-based semantic segmentation model, from Yandex researchers, proposed in March of 2022. Source: https://arxiv.org/pdf/2112.03126.pdf

Though diffusion-based semantic segmentation may currently be unsuitable for real-time applications, it would appear to have notable potential in providing better and more granular groupings of pixels during dataset pre-processing, and even in the annotation and labeling process itself; and this could be a huge aid in fighting the entanglement that can occur when generative systems struggle to separate labels from their context and environments.



DAMM's re-imagined semantic segmentation in action. Source: https://github.com/castorini/daam

Semantic Segmentation as a 'Guideline' in Image Synthesis

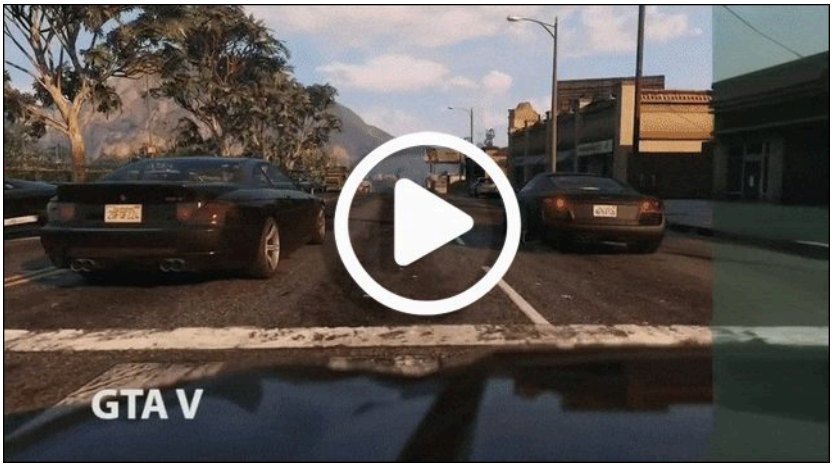
The new paradigm of ‘sketch-to-image’ applications in image synthesis over the last few years has shown the extent to which reversing the typical semantic segmentation workflow (i.e., creating delineations from static pixels) can be powerfully reversed, so that ‘imagined’ semantic masks can be in-filled by trained generative systems that can associate a label with a tranche of color, allowing the user to effectively ‘paint’ hyper-real imagery:

CLICK TO VIEW ANIMATED GIF

NVIDIA GauGAN was one of the earliest mask-based sketch-to-image systems, and part of a series of the company's explorations of the mask>real application paradigm. Here we see the user drawing segmentation maps, and the trained neural network reinterpreting the crude daubs into photorealistic landscapes and sub-facets, based on the real-world trained data. Source: https://www.youtube.com/watch?v=OGGjXG562WU

This base concept of color-coded, class-related maps as artistic ‘guidelines’ has gained much wider adoption since those early experiments with landscape-based synthesis.

In 2021, researchers from Intel developed an impressive system of neural rendering, whereby segmented classes were derived from original low-quality game footage and ‘hyper-scaled’ up not only in resolution, but from CGI to photo-real appearance trained on real-world imagery from Mapillary:



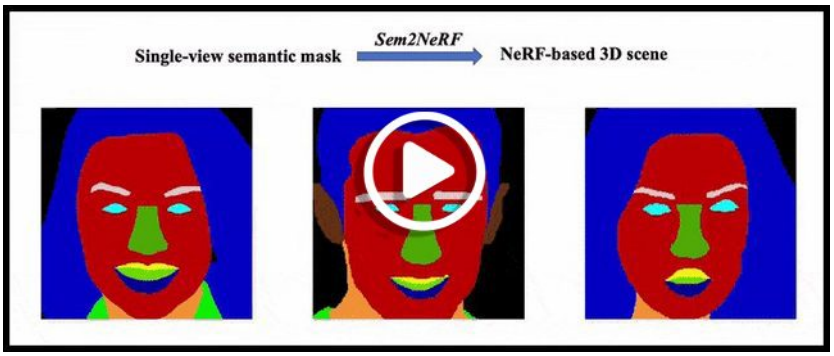
CLICK TO VIEW ANIMATED GIF

Intel's 2021 neural rendering system in action. Source: https://www.youtube.com/watch?v=P1IcaBn3ej0

Here semantic segmentation is acting as an interstitial interpretation layer, converting the rasterized game footage into vector-based segmentation labels, which are then passed to a network that reinterprets them based on the trained data.

Likewise for facial synthesis, semantic segmentation maps are being actively used as areas for interpretation by networks trained on real data. In the 2022 Sem2NeRF image translation system, free viewpoint image generation into Neural Radiance Fields (NeRF), is facilitated by semantic segmentation masks that condition the neural representation:

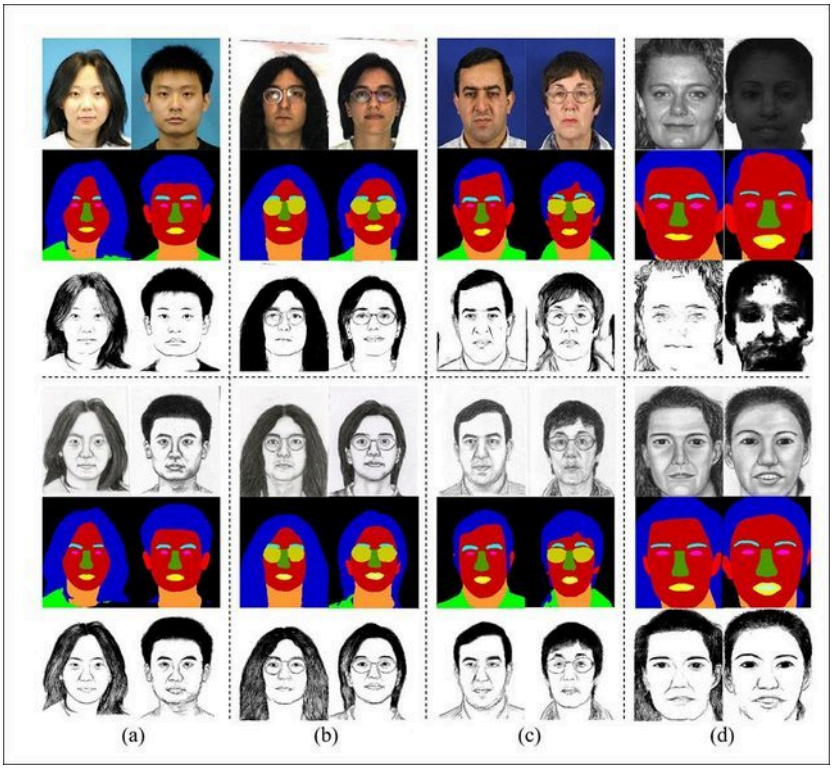
CONTINUED ON NEXT PAGE



CLICK TO VIEW ANIMATED GIF

Sem2NeRF generates faces trained on real data, using semantic segmentation maps as a means of instrumentality. Source: https://www.youtube.com/watch?v=cYr3Dz8N_9E

A 2019 outing from Northeastern University posited the use of semantic segmentation as a guideline for GAN-based facial synthesis, while a smorgasbord of other applications are researching the possibilities of turning photos into sketches using the reductionism of segmentation maps to clear out the confusion, clutter and entanglement that this task has traditionally entailed.



From the paper ' Biphasic Face Photo-Sketch Synthesis via Semantic-Driven Generative Adversarial Network with Graph Representation Learning', semantic segmentation masks are used to generate non-photoreal faces; this time the generative trained dataset has made use of real-world artistic and interpretive images. Source: https://arxiv.org/pdf/2201.01592.pdf

Conclusion

Though semantic segmentation was originally intended for more prosaic pursuits, such as applications in security, robotics, and medicine, its capacity to capture the essential space of a distinct entity (such as a 'person' or a 'cat') has renewed value in the world of multimodal image synthesis, where labels are no longer 'disposable' orientation tools to calibrate training routines, but rather active and essential assets of the generative process.