A model that can realistically insert humans into images

May 24 2023, by Ingrid Fadelli

Demonstrating the model's capability to realistically insert individuals into diverse scenes. The challenges include inferring a plausible pose given the scene context, re-posing the person, and harmonizing the insertion with respect to lighting and shadows. Credit: Kulal et al

The recent advent of generative models, computational tools that can generate new texts or images based on the data they are trained on, opened interesting new possibilities for the creative industries. For example, they allow artists and digital content creators to easily produce realistic media content that integrates elements of different images or videos.

Inspired by these recent advances, researchers at Stanford University, UC Berkeley and Adobe Research have developed a new model that can realistically insert specific humans into different scenes, for instance showing them as they exercise in the gym, watch a sunset on the beach, and so on.

Their proposed architecture, which is based on a class of generative models known as diffusion models, was introduced in a paper pre-published on the *arXiv* server and set to be presented at the Conference on Computer Vision and Pattern Recognition (CVPR) 2023 in Vancouver this June.

"Visual systems inherently possess the ability to infer potential actions or interactions that an environment or a scene allows, a concept known as 'affordances,'" Sumith Kulal, one of the researchers who carried out the study, told Tech Xplore.

"This has been the subject of extensive research within the fields of vision, psychology and cognitive sciences. Computational models for affordance perception developed over the last two decades were often constrained due to inherent limitations in their methodologies and datasets. However, the impressive realism demonstrated by large-scale generative models showed a promising avenue for progress. With these insights, we aimed to craft a model that could explicitly tease out these affordances."

The primary objective of the study by Kulal and his colleagues was to apply generative models to the task of affordance perception, in the hope of achieving more reliable and realistic results. In their recent paper, they specifically focused on the problem of realistically inserting a person into a given scene.

Showcasing the model's auxiliary tasks at inference time, which include hallucinating a person compatible with the scene, generating a scene suitable for a given person, and swapping clothes in a virtual try-on setting. Credit: Kulal et al

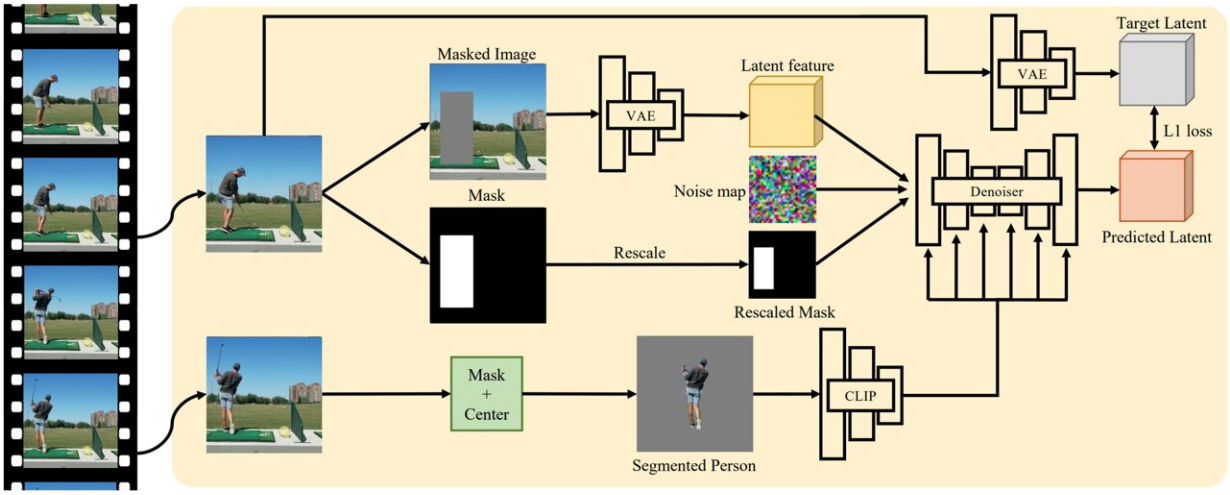
"Our inputs include an image of a person and a scene image with a designated region, and the output is a realistic scene image that now includes the person," Kulal explained. "Our large-scale generative model, trained on a dataset comprised of millions of videos, offers greater generalization to novel scenes and people. Moreover, our model exhibits a range of intriguing auxiliary capabilities such as person hallucination and virtual try-on."

Kulal and his colleagues trained a diffusion model, a type of generative model that can turn "noise" into a desired image, using a self-supervised training approach. Diffusion models essentially work by "destroying" the data they are trained on, adding "noise" to it and then recovering some of the original data by reversing this process.

During training, the model created by the researchers was fed videos

showing a human being moving within a given scene, and it randomly selected two frames from each of these videos. The humans in the first frame are masked, meaning that a region of pixels around the human is grayed out.

The model then tries to reconstruct individuals in this masked frame using the same, unmasked, individuals in the second frame as a conditioning signal. Over time, the model can thus learn to realistically replicate how humans would look if they were placed in specific scenes.



Self-supervising training scheme. Two random frames are extracted, with the person in the first frame being masked out. The person from the second frame is then utilized as a conditioning element to inpaint the image. Credit: Kulal et al

"Our method compels the model to deduce a possible pose from the scene context, re-pose the person, and harmonize the insertion," Kulal said. "A key ingredient of this approach is our dataset, composed of millions of human videos. Due to its scale, our model, similar in architecture to the Stable Diffusion model, generalizes exceptionally well to diverse inputs."

The researchers evaluated their generative model in a series of preliminary tests, where they fed it new images of people and scenes, to then observe how well it placed these people in the scenes. They found that it performed remarkably well, creating edited images that looked quite realistic. The affordances predicted by their model are better and work in a more diverse setting than those produced by non-generative models introduced in the past.

"We were thrilled to observe the model's effectiveness for a broad range of scene and person images, accurately identifying the appropriate affordances in most instances," Kulal said. "We anticipate that our findings will significantly contribute to future research in affordance perception and related areas. The implications for robotics research, where identifying potential interaction opportunities is crucial, are also substantial. Furthermore, our model has practical applications in creating realistic media (such as images and videos)."

In the future, the model developed by Kulal and his colleagues could be integrated within a number of creative software tools to broaden their image editing functionalities, ultimately supporting the work of artists and media creators. It could also be added to photo editing smartphone applications, allowing users to easily and realistically insert a person in photographs.

"This work offers several potential avenues for future exploration," Kulal added. "We are considering incorporating greater controllability into the generated pose, with recent works like ControlNet providing relevant insights. One could also expand this system to generate realistic videos of humans moving within scenes, as opposed to static images. We are also interested in model efficiency, questioning whether we can achieve the same quality with a smaller, faster model. Finally, the methods presented in this paper aren't restricted to humans; we could generalize this approach to all objects."

More information: Sumith Kulal et al, Putting People in Their Place: Affordance-Aware Human Insertion into Scenes, *arXiv* (2023). DOI: 10.48550/arxiv.2304.14406