Augmented Reality World Editor

Jake Guida
University of California Santa Barbara
jlguida@ucsb.edu

Misha Sra
University of California Santa Barbara
sra@ucsb.edu

ABSTRACT

Image inpainting allows for filling masked areas of an image with synthesized content that is indistinguishable from its environment. We present a video inpainting pipeline that enables users to "erase" physical objects in their environment using a mobile device. The pipeline includes an augmented reality application and an on-device conditional adversarial model for generating the inpainted textures. Users are able to interactively remove clutter in their physical space in realtime. The pipeline preserves frame to frame coherence, even with camera movements, using the Google ARCore SDK.

CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

KEYWORDS

Augmented Reality, Video Inpainting, Conditional GANs

ACM Reference Format:
Jake Guida and Misha Sra. 2020. Augmented Reality World Editor. In *26th ACM Symposium on Virtual Reality Software and Technology (VRST '20), November 1–4, 2020, Virtual Event, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3385956.3422125

1 INTRODUCTION

Augmented Reality (AR) makes it possible to visualize digital elements overlaid onto the physical environment. The power of AR lies in using detected features of the environment to place virtual objects with shadows, physics and occlusion to make them seem part of the real world. However, if a virtual object partially overlaps a physical object, the illusion of realism is lost [3]. Therefore, it seems reasonable to consider removing real objects to help enhance the AR illusion. The approach of removing objects from a video stream is described as *diminished reality* and has been explored in prior work using a combination of image processing and computer vision techniques [3]. Our vision is to provide an easy to use AR application that can help users remove undesirable real world objects visible in their smartphone camera's field of view, whether their end goal is to add new virtual objects to the scene or to simply take photos without clutter. We see this as effectively "Photoshopping" the real world before taking a photo instead of after, as is usual. We demonstrate our vision through a fully implemented mobile AR application that allows users to virtually remove objects from wooden surfaces (Figure 2). Diminished reality means virtually removing something from the camera's view. The object to be removed needs to be detected and the corresponding object area needs to be filled in with the background texture [3]. This filling-in is called image inpainting or context-aware fill (available in Photoshop) where rendering a new texture on top of the original static image hides the unwanted object. Video inpainting, on the other hand, has been used in video post-production for removal of undesired objects and for video frame repair when digitizing vintage movies [1]. In realtime video inpainting, the unwanted objects need to be tracked and removed by rendering the new texture in every frame. While a few different approaches to realtime or offline video inpainting like exemplar-based [1] or temporal information-based methods have been developed, in this work we explore using a generative model running in realtime on the user's smartphone. To use our app on an AR HMD, the interface design will need to be significantly changed to accommodate gesture or handheld controller input vs. the currently used touch-based input for mobile AR.

VRST '20, November 1–4, 2020, Virtual Event, Canada

ACM ISBN 978-1-4503-7619-8/20/11.
https://doi.org/10.1145/3385956.3422125

2 SYSTEM OVERVIEW

Image inpainting is an important task in computer vision with applications in photo editing, image-based rendering, and computational photography. We present a complete pipeline, including a trained model and an AR application for realtime video inpainting on a mobile device with a single RGB camera. Our goal is to interactively enable believable clutter removal. Our system is composed of two main components: 1) the AR application, and 2) a trained generative adversarial network or GAN. We used Google's ARCore and Sceneform SDKs to build the AR application. OpenCV was used for both augmenting the training data, and for processing the input data for the on-device TFLite model to generate **inference**. We used the Pix2Pix [2] conditional generative model and trained it on our wood texture dataset using Keras. While GANs learn a mapping from a random noise vector z to output image y, $G : z \rightarrow y$, conditional GANs learn a mapping from an image x and random noise vector z, to y, $G : x, z \rightarrow y$ [2]. The Pix2Pix model [2] in our pipeline does not use a random noise vector as part of its input.

2.1 Mobile AR Application

We start by asking the user to scan their physical environment, especially horizontal surfaces like the floor and tabletops. Once the space is scanned, the user is asked to drag a white marker over the objects they would like to remove. This marker is anchored in 3D space and occludes the objects on the floor from all angles even as the user moves their camera or walks around. The user can rotate and resize the marker until they are satisfied that it completely masks the physical objects before starting the inpainting process. To begin realtime inpainting the user clicks on a floating action button which sends the masked input to the trained model running on-device. Using that input, the model outputs a prediction that