A machine learning system that is capable of virtually removing buildings from a live view
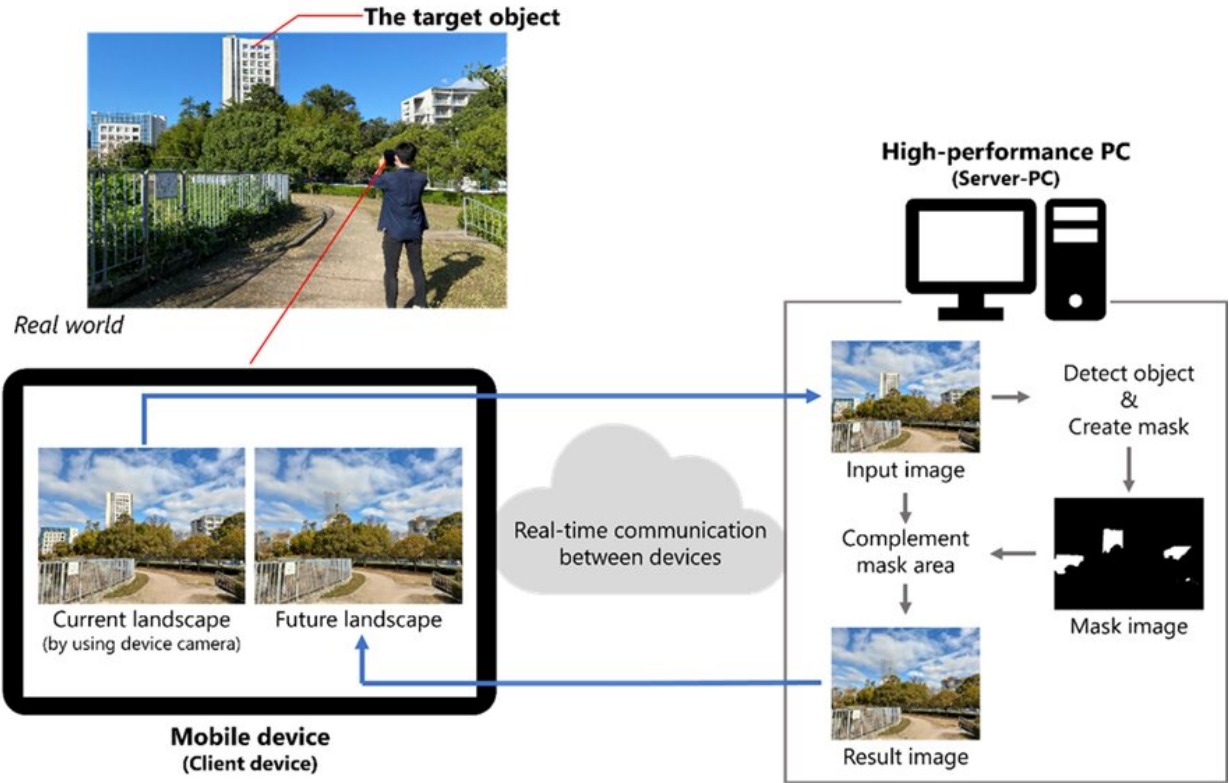
August 3 2022

Fig.1. Overview of the proposed method. An image of the current landscape is acquired by the mobile terminal and sent to the server PC. The server detects the target building and generates a mask. The area to be complemented is set from the mask image, and the input image is automatically altered based on the features around the target area. The output image based on the digital completion is sent to the mobile terminal as a future landscape after demolition to be displayed on the DR display. Credit: Takuya Kikuchi et al.

Scientists at Osaka University have created a machine learning system that is capable of virtually removing buildings from a live view. By using generative adversarial networks (GAN) algorithms running on a remote server, the team was able to stream in real-time on a mobile device. This work can help accelerate the process of urban renewal based on community agreement.

Some necessary urban renewal tasks, such as demolishing old buildings, are delayed due to the difficulty in convincing stakeholders to commit resources to a project. For instance, differences in understanding about the plan among building owners and nearby residents may lead to conflict and delays. This may result in a paradox in which tasks would be feasible to begin only after they are already accomplished. Without access to a time machine, this may seem to lead to intractable situations in civil planning.

Now, a team of researchers at Osaka University have help to address this concern in the form of a new algorithm based on machine learning that provides augmented reality real-time video demonstrating the view after a building is removed. "Our method enables users to intuitively understand what the future landscape will look like, which can contribute to reducing the time and cost for forming a consensus," first author Takuya Kikuchi says. Communication between a mobile device and a server means that all the processing can be done remotely, so any smart phone or tablet can be used at the location of the building. To speed up the algorithm so it can provide real-time augmented video, the team used semantic segmentation on the input image. This allows the deep learning model to classify images pixel by pixel, as opposed to conventional methods that try to perform 3D object detection.
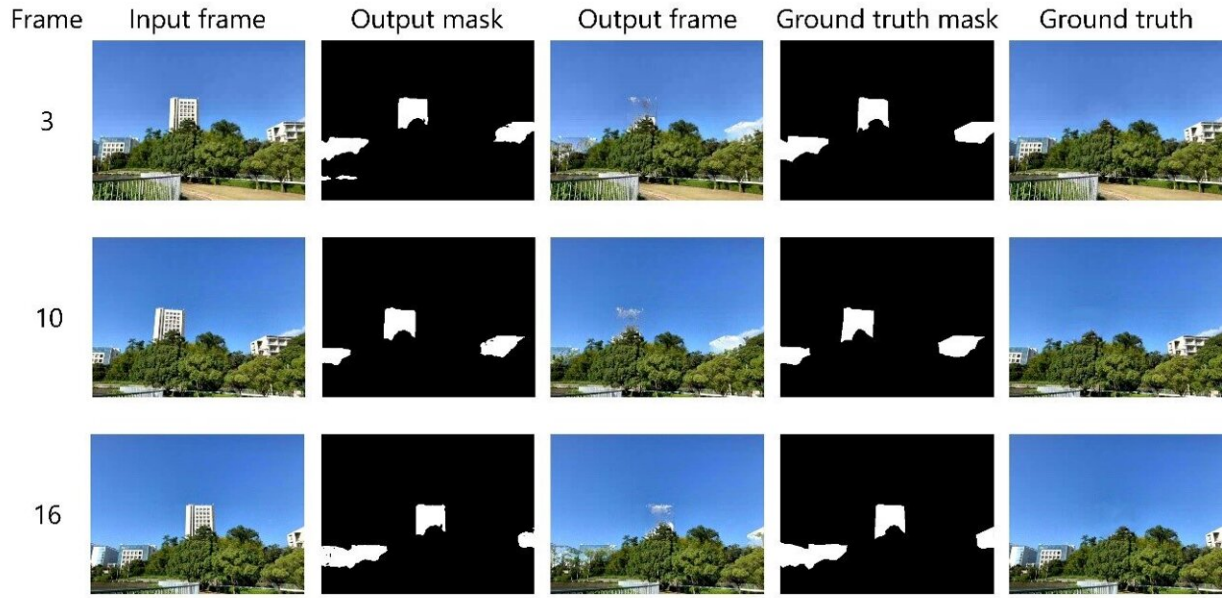
Fig.2. A future landscape after demolition visualized by the implemented DR system (Output frame). Input frame: Input image, which is the current landscape. Output mask: Result of automatic building detection and masking. Output frame: Result of automatic completion of the building area by GAN. Ground truth mask: Correct image for mask. Ground truth: Correct image for output frame. Credit: Takuya Kikuchi et al.

GAN algorithms use two competing neural networks, a generator and a discriminator. The generator is trained to create increasingly realistic images, while the discriminator is tasked with distinguishing if the image was real or artificially generated. "By learning in this way, the GAN algorithm can produce images that do not actually exist but are plausible," corresponding author Tomohiro Fukuda says. In this case, high accuracy processing was possible as long as the building to be removed from the landscape did not take up more than 15% of the screen. On the basis of field tests, the team was able to achieve virtual demolition video to be streamed at an average rate of 5.71 frames per

second, which may greatly assist in on-site community enhancement.

| Background Mask ratio | Truth | Output by GAN | | $\Delta E^*_{00} \leq 6.5$ | $\Delta E^*_{00} \leq 13.0$ |
		Using GSV	Using ImageNet	GSV | ImageNet	GSV | ImageNet
a sky 7.6%				64.5% | 60.9%	89.0% | 87.8%
b tree 10.1%				21.6% | 23.1%	56.6% | 62.9%
c tree, sky, object 14.9%				29.6% | 32.0%	73.1% | 62.0%
d tree, sky 9.1%				32.3% | 36.3%	54.9% | 74.3%

Comparison of the results of completion with GAN using two different datasets, Google Street View (GSV) and ImageNet, along with the correct image. An example of showing a comparison of completion accuracy. This is based on the size of the background element and the completed area, and the completion accuracy, shown as the difference in color. The degree to which completion accuracy, evaluated as the percentage of CIEDE2000 below a threshold value, varies with the size of background elements and completion regions, and the type of training dataset, was analyzed. Credit: Takuya Kikuchi et al.

More information: Takuya Kikuchi et al, Diminished reality using semantic segmentation and generative adversarial network for landscape assessment: Evaluation of image inpainting according to colour vision, *Journal of Computational Design and Engineering* (2022). [DOI:](#)