IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 20, 2021, accepted August 11, 2021, date of publication August 19, 2021, date of current version August 30, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3106124

Automatic Object Removal With Obstructed Façades Completion Using Semantic Segmentation and Generative Adversarial Inpainting

JIAXIN ZHANG, TOMOHIRO FUKUDA, AND NOBUYOSHI YABUKI, (Member, IEEE)
Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, Suita, Osaka 565-0871, Japan

Corresponding author: Tomohiro Fukuda (fukuda@see.eng.osaka-u.ac.jp)

This work was supported in part by Japan Society for the Promotion of Science (JSPS) Grants-in-Aid for Scientific Research (KAKENHI) under Grant JP19K12681.

ABSTRACT Automatic object removal with obstructed façades completion in the urban environment is essential for many applications such as scene restoration, environmental impact assessment, and urban mapping. However, the previous object removal typically requires a user to manually create a mask around unwanted objects and obtain background façade information in advance, which would be labor-intensive when implementing multitasking projects. Moreover, accurately detecting objects to be removed in the cityscape and inpainting the static obstructed building façade to obtain plausible images are the main challenges for this objective. To overcome these difficulties, this study addresses the object removal with the façade inpainting problem from the following two aspects. First, we proposed an image-based cityscape elimination method for automatic object removal and façade inpainting by applying semantic segmentation to detect several classes, including pedestrians, riders, vegetation, and cars, as well as using generative adversarial networks (GANs) for filling detected regions by background textures and patching information from street-level imagery. Second, we proposed a workflow to filter unoccluded building façades from street view images automatically and tailored a dataset for the GAN-based image inpainting model with original and mask images. Furthermore, several full-reference image quality assessment (IQA) metrics are introduced to evaluate the generated image quality. Validation results demonstrated the feasibility and effectiveness of our proposed method, and the synthetic image is visually realistic and semantically consistent.

INDEX TERMS Generative adversarial networks, semantic segmentation, automatic object removal, façade inpainting, street view images.

I. INTRODUCTION

Automatic object removal is a widely studied and fundamental task for environmental impact assessment due to the sheer number of unwanted objects (e.g., pedestrians, riders, vegetation, and cars) that frequently occlude the scene hinders significant tasks such as stakeholder engagement [1] and design support [2]. By visually removing unwanted visual elements in redevelopment projects, stakeholders, including experts such as construction managers and architects (decision-makers), as well as non-experts such as residents, can build a consensus of what the environment will look like after removal and evaluate the impact on the surroundings in concrete terms. In design support study, object removal techniques combined with augmented reality (AR) can address the collision problem between planned design objects and existing objects. AR allows users to have an immersive view of 3D design models superimposed on the natural world [3]. However, suppose that old structures planned for demolition are still exist in the project while AR is simulated. In that case, the newly designed 3D virtual objects will be intermingled with the existing ones, producing an inaccurate visualization. By eliminating and adding objects virtually in a perceived environment, the stakeholders can assess the future urban environment design [4].

Several studies have been made to remove objects automatically in the urban environment, from filtering out regions with unwanted objects [5] to assuming a static

The associate editor coordinating the review of this manuscript and approving it for publication was Tai-Hoon Kim.

scene and classifying object regions as outliers [6]. Recently, learning-based methods for background texture inpainting have yielded promising results [7], [8]. These methods first detect regions containing unwanted objects at the pixel level using semantic segmentation [9], then synthesize the context in those regions using image inpainting techniques. Thus, the main challenges for our objective are accurately detecting objects to be removed in the urban scenes and inpainting the statically occluded backgrounds to get plausible images.

With the development of Deep Convolutional Neural Networks (DCNNs) and semantic segmentation datasets for urban driving scenes, significant progress has been achieved in the automatic segmentation of street elements, where unwanted objects can be detected from street-level images with high accuracy by creating target object masks [10], [11]. In addition, there are many applications of image inpainting in urban scene completion. For example, the entire building can be eliminated by filling the background sky for environmental assessment [12], synthesizing the façade texture during the building renovation and digital heritage restoration [13], and virtually demolishing the landscape in redevelopment [14]. These studies match and copy background patches to fill in missing images with good results in image restoration tasks. Still, they easily lead to failures in complicated and non-repeatable scenes [15]. Visual authenticity and semantic correctness are essential indicators for evaluating the quality of synthetic images. Recent encouraging advances in data-driven image inpainting methods, which are more effective than classical methods in handling object removal from complex scenes and large occlusion rate images, have attracted the interest of researchers [16].

This study aims to develop an image-based cityscape elimination method for automatic object removal and façade inpainting. For the specific task of façade inpainting in the street-level scenes, we adopt an end-to-end deep learning-based image inpainting model to fill the detected regions by training customized datasets instead of using open-source datasets (e.g., Place2, Cityscapes, and ImageNet). Firstly, we collected street networks within the Kansai region, Japan, and downloaded street view images through an open web API. While these images had a large number of undesirable data, such as occluded building façade and urban landscape, we compared several state-of-the-art DCNNs for image classification to clean up invalid data in street view images and subsequently built a dataset consisting of building façades for learning-based image inpainting. Then, semantic segmentation was used to detect regions of unwanted objects automatically. Next, a GAN-based image inpainting method was proposed to provide a cost-effective tool for matching the physical space with the digital objects in large-scale images by filling the missing region contents of building façades with contextual attention. Furthermore, the qualitative and quantitative validations were presented for assessing the quality of generated imagery.

The contributions of the paper are summarized as follows:

- A general framework for façade inpainting and automatic object removal using object detection and image inpainting was proposed, and the object classes include pedestrians, riders, vegetation, and cars.
- A street view benchmark dataset was built to classify building façades without obstacles from street-level imagery, and a dataset was tailored for GAN-based façade inpainting in urban scenes with unoccluded façade images, mask images, and semantic segmentation labels.
- The proposed method is more efficient than previous methods when dealing with multitasking projects where background information cannot be obtained in advance.

II. RELATED WORK

A. OBJECT DETECTION AND REMOVAL

In the process of removing, eliminating, or reducing external perceptible stimuli, it is necessary to detect unwanted areas first, called region of interest (ROI) detection [17]. Then remove the detected region and fill in the background texture. The recent emergence of object detection based on deep learning has shown a powerful performance for ROI detection. CNN is a multi-layer neural network designed to recognize pixel visual patterns directly, and many robust CNN architectures have been developed, such as AlexNet [18], and ResNet [19], and Xception [20]. Along with the development of CNN, numerous CNN-based object detection algorithms have been developed, such as You Only Look Once (YOLO) [21] and DeepLab v3+ [22], have been proposed. In these methods, objects can be detected and segmented according to their contours. However, when using CNN-based object detection to determine the ROI, the ROI need not be a contour of the target object but rather a mask that wholly covers the target object. The ROI needs to be filled after object detection, and there are two main methods: observation and inpainting. Observation requires a pre-captured image of the background scene [4], which can be used as a reference to replace foreground obstacles directly. The other method is inpainting, which uses texture and patch information from the source image to fill the detected area [23]. However, it is hard to obtain background pictures when obstacles cannot be moved in practical projects. Therefore, the inpainting method without pre-processing is more feasible than the observation method in handling object removal in street-level scenes.

B. GENERATIVE ADVERSARIAL INPAINTING

Previous image inpainting methods, such as exemplar-based inpainting, can find the approximate nearest neighbor to match image patches [16]. With the rise of learning-based image inpainting methods, the inpainting results have become more realistic and reasonable [24]. In general, the existing methods for image hole inpainting can be divided into three categories:

- *Inpainting by Replication:* These methods tried to explicitly borrow content or texture from the

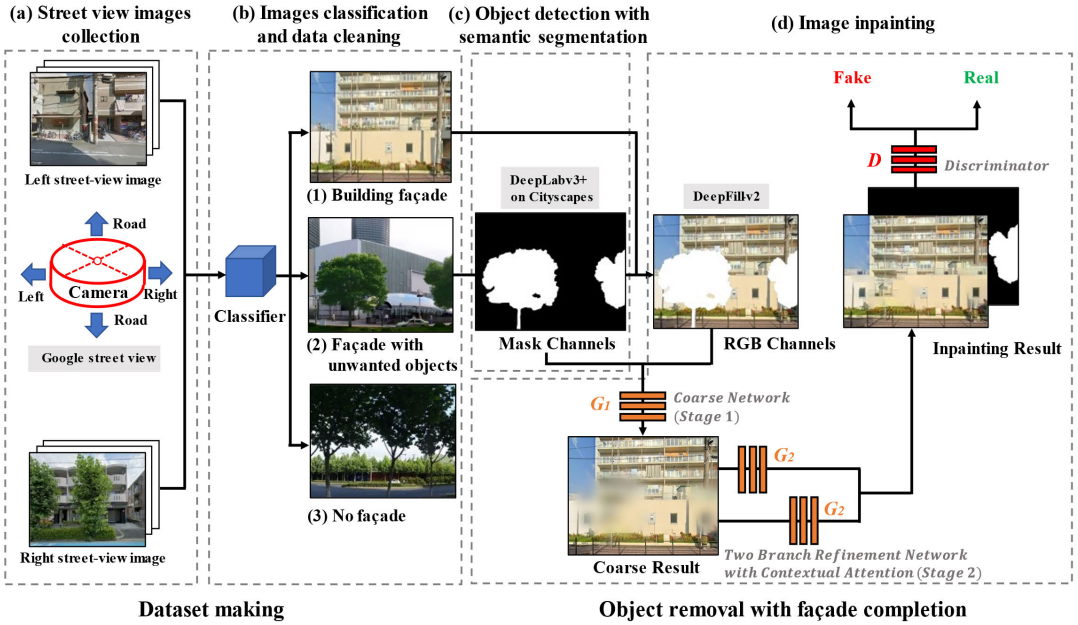

FIGURE 1. The overall workflow of the proposed method.

surrounding environment to fill the missing areas. An example of unsupervised learning was a context copy method using surrounding image information to predict the loss of contents [25], but image replication tends to fail when dealing with complex scenes.

- *Inpainting by Modeling:* These methods used large external databases to imagine missing pixels in a data-driven manner. They tried to learn to model the distribution of the training images and assumed that regions surrounded by similar backgrounds might have comparable content [26]. These methods can effectively find sample images with sufficient visual similarity to the query, but they were easy to fail with no similar examples in the database.
- *Combining the Two:* The third class of approaches attempted to combine the two to overcome the limitations of replication methods or modeling methods. Not only do these methods learned to build image distributions in a data-driven manner, but they were also designed to explicitly borrow patches or features from background regions [16]. However, when the training dataset and the content of the processed images do not match, the generated image quality is not satisfactory. Image inpainting works better by customizing the dataset rather than using a generic dataset for a specific task.

This paper introduces an end-to-end image inpainting model combining replication and modeling proposed by Yu *et al.*, named DeepFill-v2 [8]. The image inpainting model adopts stacked generative networks to ensure that the color and texture of generated areas are consistent with their surroundings. Moreover, the contextual attention model is integrated into networks, enabling the borrowing of detailed data from distant spatial locations. To overcome the mismatch between the opensource dataset and the content of the façade inpainting task. This approach focuses on training our tailored dataset collected from street view images.

C. STREET VIEW IMAGES

Street view images are among the most widely used City Scene Data groups and have been extensively used for autonomous driving [27] and urban environment analysis [28], [29]. Multiple databases have been established, and the frequently used data sources are Google Street View (GSV), Baidu Street View (BSV), and Tencent Street View (TSV) [10]. These street view data are open-source, easy-to-obtain, and cover most streets in the world. From the perspective of pedestrians and cars, the system records the street scene in detail. The street view image can describe the fine-grained physical environment elements, including urban infrastructure, human-made cityscapes, and natural landscapes. Street view images will provide sufficient façade images to ensure reliable results for the GANs model training [30]. However, the street view database does not directly give the orthographic projection of building façade pictures. To solve this issue, it is necessary to obtain the spatial coordinate information of each road network and to bring the vertical angle of the road into the data acquisition of the street view service to get images perpendicular to the building façade [31].

D. GENERATED IMAGE QUALITY ASSESSMENT

Despite the remarkable achievements in GAN-based image restoration, there were many challenges of quantitative evaluation metrics that satisfy human visual perception while using objective computational values [32]. Most GAN-based image generation methods concentrated on synthesizing richer

textures, typically evaluated by measuring the similarity between generated images and ground-truth through full-reference metrics such as mean squared error (MSE), peak signal-to-noise ratio (PSNR), structural similarity (SSIM) index, and feature similarity indexing method (FSIM) [33]. The previous study in high-resolution image generation indicated that the information fidelity criterion (IFC) performs favorably among full-reference metrics for image quality assessment (IQA) [34]. However, generated image assessment using fully referenced indicators required ground-truth images that were often difficult to obtain in practice. To obtain the ground-truth images, we tailored the dataset for a free-form façade inpainting GAN from the urban streetscape as the original image and superimposed the mask on the image as the input. This study introduced several full-reference metrics to evaluate the feature similarity of the generated image to the ground-truth image.

III. METHOD AND MATERIALS

Figure 1 illustrates the overall workflow of automatic object removal and façade inpainting using semantic segmentation and GAN. First, we need to make the building façade dataset for GAN-based image inpainting, and these pictures can be obtained from street view services and be cleaned using a classifier. The street-level obstacles can then be detected by a semantic segmentation algorithm based on the Cityscapes dataset. Finally, we introduce a free-form image inpainting tool to fill the blank with contextual attention.

A. STREET VIEW IMAGES COLLECTION

To create a dataset of façade inpainting with various streetscape categories, we downloaded the road networks for multiple cities using open source geographic information data [35], as shown in Figure 2**a**. The geographic coordinates of road points were generated by equally sampling direction for each sampling point from the Google Map API (viewing angle is 90 degrees, the horizontal angle is 0 degrees, the compass heading of the camera is θ, and the picture size is 680 × 512 pixels). Finally, as shown in Figure 2**b** and 2**c**, to ensure that the angle of the crawled picture is perpendicular to the street, θ is calculated as follows:

$$\theta = arctan\,(y_A - y_B, x_A - x_B) \tag{1}$$

where point A (x_A, y_A) and Point B (x_B, y_B) are two adjacent points on the road centerline, and the angle θ is the deflection angle that grabs the orthographic projection of the building façade in the online street view service. The existing building façade in the urban environment can be obtained (Figure 2**d**), and these images can be made into the training set of the image generative inpainting network. Figure 3 demonstrates the structure of the street view image recording.

B. DATASET

Recent research indicates that increasing the amount of valid training data contributes to remarkable improvement in CNNs [36]. Therefore, we need to get more valuable data

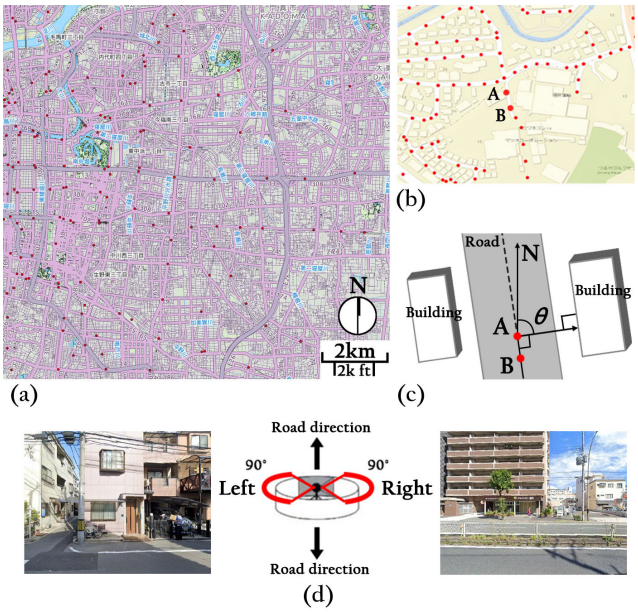

FIGURE 2. The collection method of the perpendicular street facade.

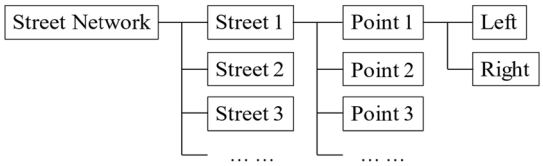

FIGURE 3. The structure of data recording.

TABLE 1. The types of data augmentation: image normalization, shift, scaling, and horizontal flip.

Data augmentation	Value
Image normalization	1/255
Width shift	0.1
Height shift	0.1
Size scaling	[0.9, 1.2]
Horizontal mirroring	-

by customizing the dataset. The image generative inpainting model requires learning textures from a large number of unobscured façades images. Since there are many noise and undesirable pictures in the collected street façade pictures, we need to use an image classification method to clean these data. To create the image classifier, we first create a corresponding streetscape benchmark dataset from three categories (i.e., building façade, façade with unwanted objects, and no façade), as shown in Figure 1**b**. 2700 street façade images are manually selected from street view services, with 900 images in each class. Data augmentation is then used to expand the training sample while increasing its diversity, which helps avoid overfitting and improves model performance [37]. Table 1 shows the geometric transformations for data augmentation, including image normalization, width shift, height shift, size scaling, and horizontal mirroring. Finally, we compare several CNN-based architectures and choose the best performance trained model to classify the building façade images from the street view images as the

FIGURE 4. Labels can be detected and removed in this method: Pedestrian, rider, vegetation, and car semantic segmentation using DeepLab v3+ on Cityscapes test set.

TABLE 2. Apparatus.

Content	Appellation
CPU	Intel Core i7-9700 @ 3.00GHz
GPU	Nvidia GeForce GTX 1080 Ti 11GB × 2
RAM	DDR4-2666 16GB × 2
OS	Window 10 Home 64bit

dataset for the façade inpainting GAN, and named 'Street view dataset for building façade inpainting (SVBFI)'.

C. SEMANTIC SEGMENTATION

Semantic segmentation combines image classification, image detection and can perform categorization and annotation in terms of pixel-by-pixel in an image [38]. Semantic segmentation tasks are composed of two components: the dataset and the segmentation algorithm. In the study, unwanted objects were determined in the object segmentation dataset using Cityscapes [10]. DeepLab v3+ [39] was used for semantic segmentation.

The Cityscapes dataset is an urban streetscape dataset taken from the pedestrian perspective with 19 categories of dense pixel annotations. We can automatically extract these 19 classes, such as pedestrian, rider, vegetation, car, etc., using semantic segmentation through the Cityscapes dataset. DCNN-based semantic segmentation is widely used to detect object types and classify each pixel of different objects. DeepLab v3+ is a CNN-based model that applies the Atrous Spatial Pyramid Pooling and decoder modules to form a fast and powerful encoder-decoder network to get highly accurate segmentation results [22].

As shown in Figure 1**c**, we use DeepLab v3+ on the Cityscapes test set for object segmentation, and its mIoU can reach 82.1%. In this paper, several classes of obstacles in the streetscapes: pedestrians, riders, vegetation, and cars are taken as the specific objects to be eliminated. Through detecting by DeepLab v3+ on the Cityscapes, they are mask images in the input image of the inpainting GAN model, as illustrated in Figure 4.

D. IMAGE INPAINTING

The façade inpainting method uses the open-source model DeepFill-v2 [8], a free-form image inpainting method with gated convolution, to generate alternative contents for blank

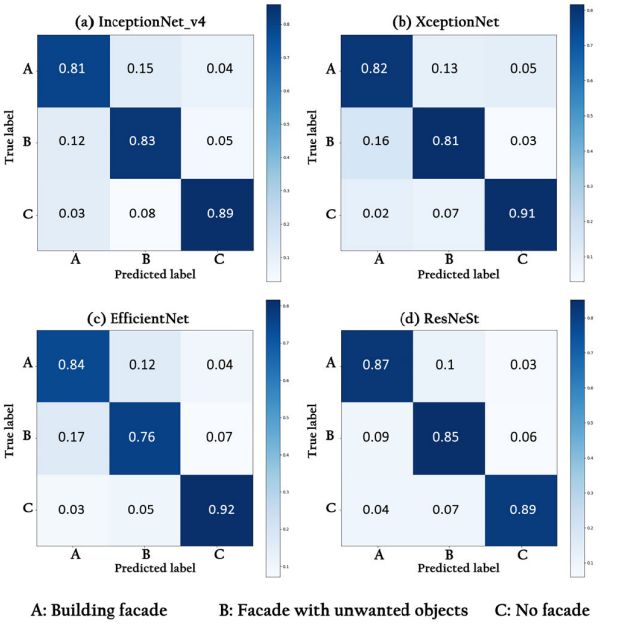

FIGURE 5. The associated normalized confusion matrices of the four networks evaluated on the test images, i.e. InceptionNet_v4 (Top-left), XceptionNet (Top-right), EfficientNet (Bottom-left) and ResNeSt (Bottom-right).

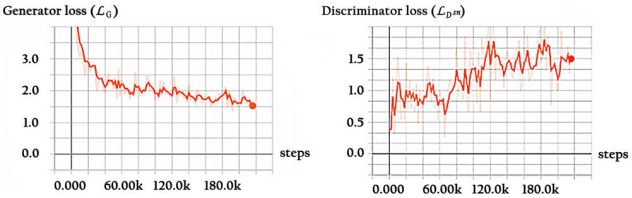

FIGURE 6. Training loss of the GAN model. Left: generator loss; Right: discriminator loss.

areas in a visually realistic and semantically correct manner. Figure 1**d** introduces the simplified overall network structure of DeepFill-v2. The input information is divided into two: RGB Channel and Mask Channel for this neural network. The model architecture includes a two-stage generator and a discriminator. The first stage of the generator is a coarse network that generates a coarse result. The second stage is a two-branch refinement network with contextual attention to form a refined result, which can significantly improve the image quality and the fidelity of the repair results. Gated convolution dramatically improves performance when the mask pictures have arbitrary shapes and the inputs are conditionally free-from, such as in the sparse sketch [8]. Thus, the mask image is composed of RGB channels with conditional inputs after detecting undesired object regions. Overall, the model can synthesize a new picture structure on the blank image in a learning-based manner while taking advantage of the surrounding picture features as a reference to generate reasonable predictions.

IV. EXPERIMENTS AND RESULTS

This section describes the SVBFI dataset making, the image inpainting model training, testing, and quantitative

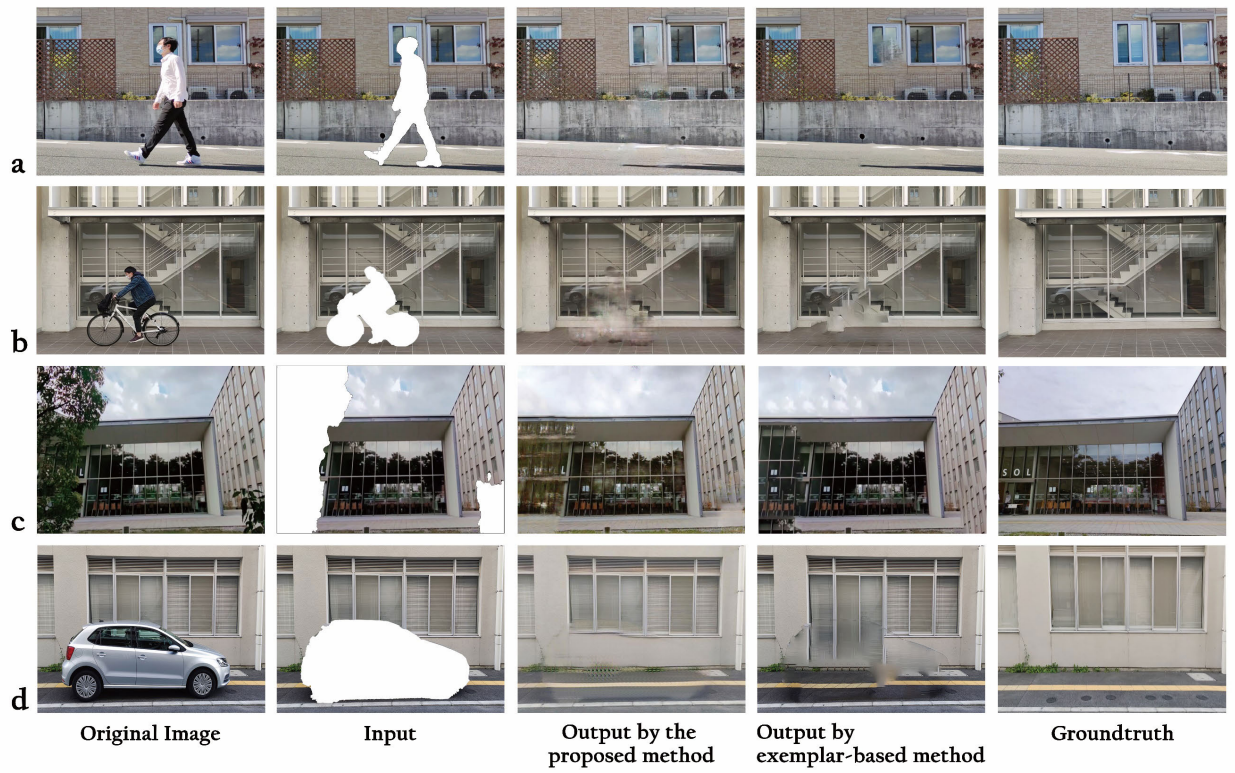

FIGURE 7. Test examples of automatic object removal and facade completion comparing with qualitative comparison. (a) People, (b) rider, (c) vegetation, (d) car.

comparisons in different scenarios. Table 2 lists the specifications of the computer was used for the experiments.

A. SVBFI DATASET MAKING

The street view images were split into a training set and testing set with 2700 pictures. 2250 images were used for training (750 images for each class), accounting for 0.83 of the entire training set. 450 testing images accounted for 0.17 of the whole training set. Several state-of-the-art CNN models are introduced, named InceptionNet_v4 [40], XceptionNet [20], EfficientNet [41], and ResNeSt [42], by fine-tuning all the convolutional layers with our benchmark dataset, and Figure 5 shows the normalized confusion matrix of the trained CNNs evaluated by our test data. The matrix value indicates the percentage that samples from one category were classified correctly into another class by the model and was a criterion to measure classification accuracy. We used F_1 score for the evaluation of model performances, which were calculated as the following equations:

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r} \tag{2}$$

where p is precision and r is recall. After calculating the F_1 scores of the four networks, the classification performance of ResNeSt performs better than the other networks. For the class of building façade, ResNeSt achieved the highest F_1 score with 0.87. Therefore, the trained ResNeSt model was chosen from Google street view for the upcoming building façade image classification.

More than 300,000 street view images were downloaded from Google street view, and the size of the pictures is 680 × 512 pixels, including unoccluded façades, façade with unwanted objects, and no façade. Filtering by the ResNeSt classier trained in the previous step, we finally got a 9,000 unoccluded façade images dataset named SVBFI. We used the SVBFI as the training set for image inpainting GAN. 80% of the images were used as the training, 10% of the images were used as the testing dataset, and 10% of the images were used for validation.

B. IMAGE INPAINTING GAN TRAINING

In general, GAN consists of a generator and a discriminator. They compete with each other so that the generated images are continuously optimized and semantically close to the ground-truth image. Recently developed spectral normalization [43] was used to stabilize the GANs training further. We utilized the SN-GANs default fast approximation algorithm for spectral normalization, and specific details were described in [43]. To discriminate if the input was real or fake, we also used the hinge loss as the objective function for the generator $\mathcal{L}_G$ and discriminator $\mathcal{L}_{D^{sn}}$.

$$\mathcal{L}_G = -\mathbb{E}_{z\sim\mathbb{P}_z(z)}[D^{sn}(G(z))] \tag{3}$$

$$\begin{aligned}\mathcal{L}_{D^{sn}} = {} & \mathbb{E}_{x\sim\mathbb{P}_{\text{data}}(x)}[\text{ReLU}\left(\mathbb{1} - D^{sn}(x)\right) \\ & + \mathbb{E}_{z\sim\mathbb{P}_z(z)}[\text{ReLU}(\mathbb{1} + D^{sn}(G(z)))]\end{aligned} \tag{4}$$

where D^{sn} represents spectral-normalized discriminator, $G(z)$ is an image inpainting network that takes incomplete image z.

In the training process, the datasets were trained for 300 epochs, which iterated 216,000 steps. Figure 6 shows the loss of generator $\mathcal{L}_G$ and discriminator $\mathcal{L}_{D^{sn}}$ in this model. The loss of the generator was decreasing, and the discriminator loss was increasing. As the generator and discriminator reach equilibrium, the overall performance of the work steadily improves.

C. TESTING AND QUALITATIVE COMPARISONS

We compare our method with the previous exemplar-based image inpainting method and the ground-truth image. Figure 7 shows testing examples of automatic object removal with façade inpainting in the street-level environment, where the object types include people, riders, greenery, and cars. The ROI for each object was covered by masks filled with contexture attention, and no post-processing was conducted to ensure fairness.

The previous studies have shown that the detailed background images are complex cases for image inpainting [23], [44]. Figure 7**d** shows that the exemplar-based method has apparent visual incongruity and incorrectly copies half of the window from the bottom. As shown in Figure 7**c**, the exemplar-based approach filled the target object area by directly borrowing the surrounding textures, resulting in the wall texture being incorrectly filled to the ground.

The DeepFill-v2 using our dataset performs visually well without noticeable color inconsistency, as shown in Figure 7a and Figure 7d, which are realistic synthetic images with seamless boundary transitions. Figure 7**b** shows that the model can fill the blank in the glass and ground of the input image, but it blurs some details. In Figure 7**c**, the proposed model can correctly identify the skyline in the input image as the actual situation. The synthesized façade and floor in the output image are visually consistent with the ground-truth image. When dealing with glass facades with complex backgrounds, the proposed method can recover coarse patterns with correct colors. The results show that the inpainting GAN method learned from massive data can effectively consider the image semantics and perform better than the exemplar-based approach in non-repeating and complicated scenes.

D. VALIDATION AND QUANTITATIVE COMPARISONS

To assess the quality of the generated images based on visual perception, two commonly used full-reference IQA metrics are introduced: PSNR and IFC. We test the proposed model on SVBFI with our validation data of 900 images and compare it with the exemplar-based approach. Unoccluded façade images are used as ground-truth, and mask images superimposed on ground-truth are used as input images. We choose the unwanted objects as mask shapes, including people, riders, trees, and cars, and place the mask area on the ground in the street-level images to simulate the position of the unwanted objects in the real scene. The masking ratio is distributed from 0 to 50% of the total image size. In addition, the validation data were divided into five intervals according to the mask area, including 0-10%, 10%-20%, 20%-30%,

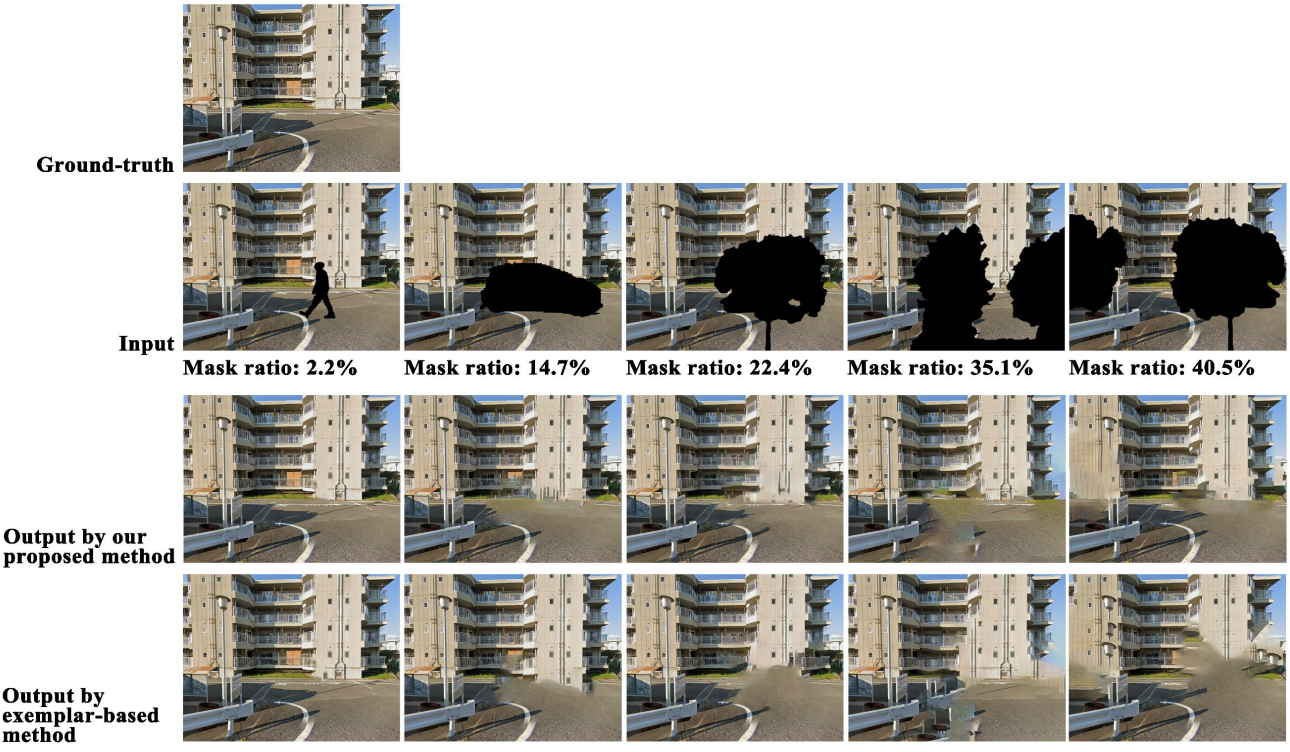

FIGURE 8. Validation examples of automatic object removal and facade completion.

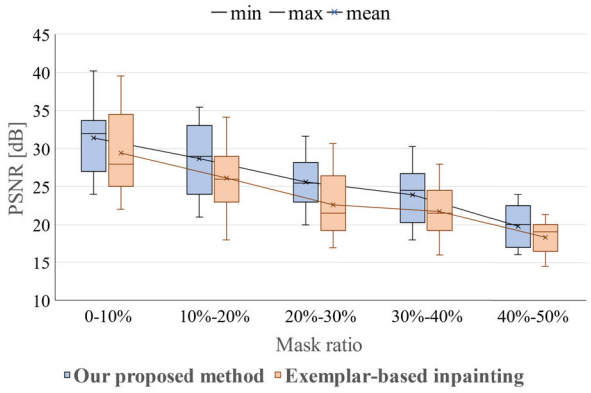

FIGURE 9. The PSNR results with our proposed method and exemplar-based method of different mask ratio on our validation data.

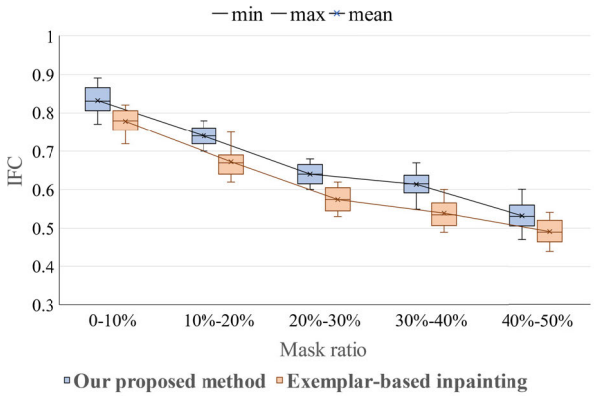

FIGURE 10. The IFC results with our proposed method and exemplar-based method of different mask ratio on our validation data.

30%-40%, 40%-50%. Figure 8 shows the validation example of our proposed method and the exemplar-based method at different mask ratios. Figure 9 and Figure 10 show the generated image quality from the PSNR and IFC results. We give the mean value of quantitative comparisons in Table 3. The quantitative comparisons of full-reference metrics indicate that the proposed method achieves better results than the exemplar-based method. Compared to the fill by replication method, our model improves PSNR by 2.26 dB and IFC 0.061 in the whole mask ratio. However, it is worth noting that the GAN-based approach using our tailor-made dataset is marginally superior to the filling through copying method for mask ratio in 40-50%, with PSNR improving by 1.56 dB and IFC improving by 0.042 in mean value. Furthermore, we calculate the processing time for the validation dataset and found that the average processing time of the exemplar-based method is 1.35 seconds per photo, while the average time of our method is 0.24 seconds.

TABLE 3. The mean value of quantitative results on test dataset with expemlar-based and our method. All mask with 0-50% area. ↑ higher is better.

Metrics	Mask	Ours	Expemlar-based method
PSNR↑	0-10%	31.39	29.43
	10-20%	28.71	26.13
	20-30%	25.64	22.61
	30-40%	23.91	21.75
	40-50%	19.87	18.31
	All	25.90	23.64
IFC↑	0-10%	0.832	0.778
	10-20%	0.743	0.673
	20-30%	0.645	0.575
	30-40%	0.612	0.538
	40-50%	0.532	0.490
	All	0.673	0.612

V. DISCUSSION

This paper described an image-based method with unwanted object removal and façade inpainting on street-level scene visualization from a human perspective. This section will discuss the advantages, potential applications, and limitations of this work.

A. ADVANTAGES

There are some advantages to this proposed method. First, our system can handle object elimination tasks for 2D images in different street scenes. In comparison to exemplar-based inpainting, our approach is suitable for completing wide holes in a shorter runtime and balances the high quality of generated images with detailed textures that have the potential for real-time applications. Second, we have established a dataset SVBFI for learning-based obstructed façade inpainting from street view images. Compared with the current open-source scene dataset, such as Places2, containing 10 million pictures and more than 400 different scene environments [8], our data is more targeted and consumes less computation for training the learning-based image inpainting task. Third, our method is more scalable than the previous onsite 3D landscape simulation [14] and does not require obtaining background façade information and a 3D model in advance. Our method can erase unnecessary elements and visualize the redevelopment project to the stakeholders by using only images. Overall, the proposed approach is efficient, lightweight, and generalized, helping assess the impact of object removal on the built environment from a human perspective.

B. POTENTIAL APPLICATIONS

These object removal tasks have numerous potential applications with obstructed façade inpainting, such as building façade synthesizing for cityscape visualization, collision problems in AR, and data augment for urban computing. This method can quickly generate 2D façade images with object removal in cityscape visualization, allowing for timely stakeholder discussions. The proposed method and dataset can be applied to digital façade generation, image restoration of urban scenes, and impact analysis after object removal in a large-scale project. Besides, the method can provide prerequisites for subsequent AR simulation designs. If AR is simulated while old structures planned for demolition are still present, our proposed method can virtually eliminate the existing objects without obtaining the background so that the new virtual elements will not be mixed with the currently unwanted objects. Furthermore, our method can also be used for data augmentation for deep learning training by removing

some visual elements in the street view without degrading their original realism. For instance, the street view dataset of cityscapes classification and segmentation can be expanded by synthesizing urban scenes that are hard-to-get or non-existent in visual perception.

C. LIMITATIONS

Despite appreciating the benefits of the study, there are still several limitations deserving further work. First, this method cannot automatically detect the shadows or remove them in subsequent paintings because it uses DeepLab v3+ based on Cityscapes without labeling the object shadows. An example is shown in the output of Figure 7**a**, where the pedestrian shadow is not removed from the frame. This deficiency can be solved by labeling the shadows of objects in the training set of the semantic segmentation model. Second, the accuracy of the ROI detection can affect the effectiveness of object elimination. If there are some errors in the mask images, it will leave the content of occlusions in the synthesized image. The mask can be larger than the object to improve the performance of the image painting. Third, this approach uses GANs to learn the background texture from large existing street view façade photos, which could fail if there are no similar façade textures in the dataset. This problem can be solved by introducing similar cityscape data from implemented projects to extend the training set.

VI. CONCLUSION

With the advancement of VR, DR, and AR technologies, the multimedia expression of urban environment visualization moves from static virtual images to dynamic scene interactions to form a paradigm shift. The visualization technology offers a communication channel for experts and non-experts, helping develop a consensus on the future urban environment design. This study proposed an image-based cityscape visualization method to automatically detect unwanted objects over the street-level scenes while removing them by inpainting the ROI without obtaining background information in advance. For this purpose, we introduce DeepLab v3+ to detect regions containing unwanted objects at the pixel level. Then we presented a general framework to tailor the SVBFI dataset for training the GAN model and use DeepFill-v2, a GAN-based image inpainting method, for obstructed façade completion. The comparison experiments show that our approach performs better than the exemplar-based inpainting method. The validation results and quantitative comparisons demonstrated the superiority of our proposed system.

By automatically removing undesired objects and inpainting obstructed building façades in a cost-efficient way, we can assess the impact of urban environmental changes on visual perception, improve information degradation during stakeholder exchanges, and encourage public engagement in discussions of the redevelopment project. In future work, we will focus on optimizing system performance at runtime with the least amount of field preparations to achieve real-time dynamic object removal. The integration of image processing techniques with artificial intelligence and mixed reality is particularly promising. We can simulate and visualize the image of a future cityscape by adding new objects or eliminating redundant ones using only images, without the need for 3D models.

ACKNOWLEDGMENT

The authors would like to thank the editors and anonymous reviewers for their constructive suggestions and comments, which helped improve this article's quality.

REFERENCES

[1] D. Kido, T. Fukuda, and N. Yabuki, "Assessing future landscapes using enhanced mixed reality with semantic segmentation by deep learning," *Adv. Eng. Informat.*, vol. 48, Apr. 2021, Art. no. 101281, doi: 10/gkgn3g.
[2] S. Siltanen, "Diminished reality for augmented reality interior design," *Vis. Comput.*, vol. 33, no. 2, pp. 193–208, 2017, doi: 10/f9m9h2.
[3] D. W. F. Van Krevelen and R. Poelman, "A survey of augmented reality technologies, applications and limitations," *Int. J. Virtual Reality*, vol. 9, no. 2, pp. 1–20, 2010, doi: 10/ggxxt5.
[4] S. Mori, S. Ikeda, and H. Saito, "A survey of diminished reality: Techniques for visually concealing, eliminating, and seeing through real objects," *IPSJ Trans. Comput. Vis. Appl.*, vol. 9, no. 1, p. 17, Jun. 2017, doi: 10.1186/s41074-017-0028-1.
[5] Y. Sun, M. Liu, and M. Q.-H. Meng, "Improving RGB-D SLAM in dynamic environments: A motion removal approach," *Robot. Auton. Syst.*, vol. 89, pp. 110–122, Mar. 2017, doi: 10/f9n37m.
[6] A. Valada, N. Radwan, and W. Burgard, "Incorporating semantic and geometric priors in deep pose regression," in *Proc. Workshop Learn. Inference Robot., Integrating Struct., Priors Models Robot., Sci. Syst. (RSS)*, vol. 1, 2018, p. 3.
[7] B. Bescos, J. Neira, R. Siegwart, and C. Cadena, "Empty cities: Image inpainting for a dynamic-object-invariant space," in *Proc. Int. Conf. Robot. Automat. (ICRA)*, 2019, pp. 5460–5466, doi: 10/gjk8dh.
[8] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.
[9] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "RGB-D object detection and semantic segmentation for autonomous manipulation in clutter," *Int. J. Robot. Res.*, vol. 37, nos. 4–5, pp. 437–451, 2018, doi: 10/gdkm7q.
[10] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223, doi: 10.1109/CVPR.2016.350.
[11] Y. Zhang, P. David, H. Foroosh, and B. Gong, "A curriculum domain adaptation approach to the semantic segmentation of urban scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1823–1841, Aug. 2019, doi: 10/ggv89g.
[12] T. Fukuda, Y. Kuwamuro, and N. Yabuki, "Optical integrity of diminished reality using deep learning," *Sharing Computable Knowl.*, vol. 1, no. 8, p. 241, 2017.
[13] D. Dai, H. Riemenschneider, G. Schmitt, and L. Van, "Example-based facade texture synthesis," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1065–1072, doi: 10/ghhnkc.
[14] D. Kido, T. Fukuda, and N. Yabuki, "Diminished reality system with real-time object detection using deep learning for onsite landscape simulation during redevelopment," *Environ. Model. Softw.*, vol. 131, Sep. 2020, Art. no. 104759.
[15] N. Zhang, H. Ji, L. Liu, and G. Wang, "Exemplar-based image inpainting using angle-aware patch matching," *EURASIP J. Image Video Process.*, vol. 2019, no. 1, pp. 1–13, 2019, doi: 10/ggb2p9.
[16] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7505–7514, doi: 10/gg9969.
[17] Y. Sun, H. Zhu, F. Zhuang, J. Gu, and Q. He, "Exploring the urban region-of-interest through the analysis of online map search queries," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2018, pp. 2269–2278, doi: 10/ghj6mp.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778, doi: 10/gdcfkn.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807, doi: 10/gfxgtm.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788, doi: 10/gc7rk9.

[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," Aug. 2018, *arXiv:1802.02611*. Accessed: Nov. 2, 2020. [Online]. Available: http://arxiv.org/abs/1802.02611

[23] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004, doi: 10/fbk8ft.

[24] Z. Qin, Q. Zeng, Y. Zong, and F. Xu, "Image inpainting based on deep learning: A review," *Displays*, vol. 69, Sep. 2021, Art. no. 102028, doi: 10/gkqdtt.

[25] T. N. Mundhenk, D. Ho, and B. Y. Chen, "Improvements to context based self-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9339–9348.

[26] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[27] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The ApolloScape dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 954–960.

[28] M. Helbich, Y. Yao, Y. Liu, J. Zhang, P. Liu, and R. Wang, "Using deep learning to examine street view green and blue spaces and their associations with geriatric depression in Beijing, China," *Environ. Int.*, vol. 126, pp. 107–117, May 2019, doi: 10/gfv446.

[29] J. Zhang, T. Fukuda, and N. Yabuki, "A large-scale measurement and quantitative analysis method of façade color in the urban street using deep learning," in *Proc. Int. Conf. Comput. Design Robot. Fabr.*, 2020, pp. 93–102.

[30] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li, "SemanticAdv: Generating adversarial examples via attribute-conditioned image editing," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 19–37.

[31] N. Ye, B. Wang, M. Kita, M. Xie, and W. Cai, "Urban commerce distribution analysis based on street view and deep learning," *IEEE Access*, vol. 7, pp. 162841–162849, 2019, doi: 10/gjjfk8.

[32] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 732–741.

[33] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study," *J. Comput. Commun.*, vol. 7, no. 3, pp. 8–18, 2019, doi: 10/ggr3gf.

[34] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005, doi: 10/dc6cn3.

[35] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, vol. 43, no. 6, pp. 32–38, 2010.

[36] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, Dec. 2018.

[37] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, 2019, doi: 10/ggb3hw.

[38] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019, doi: 10/gfwf5v.

[39] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.

[40] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 1–7.

[41] B. Koonce, "EfficientNet," in *Convolutional Neural Networks With Swift for Tensorflow*. Berlin, Germany: Springer, 2021, pp. 109–123.

[42] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, "ResNeSt: Split-attention networks," Dec. 2020, *arXiv:2004.08955*. Accessed: Jan. 25, 2021. [Online]. Available: http://arxiv.org/abs/2004.08955

[43] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957

[44] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.

JIAXIN ZHANG was born in China. He received the B.Arch. degree from Nanchang University, China, and the M.Arch. degree from Southeast University, China. He is currently pursuing the Ph.D. degree with the Division of Sustainable Energy and Environmental Engineering, Osaka University, Japan. He is also aiming at cross-boundary researches related to the urban environment and computer science. His research interests include architecture design, urban environment visualization, smart cities, and machine learning in urban data analytics.

TOMOHIRO FUKUDA is currently an Associate Professor with the Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, Japan. He majors in environmental design and information technology, before joining Osaka University. He has joined domestic and international research and design projects, such as Patio Design in Takamatsu, Kobunaki Eco village in Omihachiman, and Next Gene 21+ in Taiwan. His research interests include environmental design, development of digital design or communication tools using VR, MR, and 3DCG, and application of the system. He is a member of AIJ (Architectural Institute of Japan), JSCE (Japan Society of Civil Engineers), and CAADRIA (Computer-Aided Architectural Design Research in Asia). He was a Large Physical Model Production Member of the Aqua Metropolis Osaka for Tadao Ando Architectural Exhibition, in 2009, an Executive Board Member of the Non-Profit Organization Alternative Tourism Club, and a Tourist Guide of Osaka Community-Based Tourism.

NOBUYOSHI YABUKI (Member, IEEE) received the B.E. degree in civil engineering from The University of Tokyo, in 1982, and the M.S. and Ph.D. degrees in civil engineering from Stanford University, USA, in 1989 and 1992, respectively. He is currently a Professor with the Division of Sustainable Energy and Environmental Engineering, Graduate School of Engineering, Osaka University, Japan. He leads the Environmental Design and Information Technology Laboratory. He has authored over 100 journal and international conference papers. His research interests include developing product models, building information modeling (BIM), applying VR and AR to civil and building engineering, 4D CAD, RFID, and sensors for inspection and monitoring of structures, knowledge discovery from large amounts of sensors, and product model data. He is a member of JSCE, ASCE, ACM, AIJ, VR Society of Japan, and Japan Society of Kansei Engineering. He was a recipient of the Fulbright Scholarship for his M.S. and Ph.D. degrees. He also leads the Civil Engineering Committee of buildingSMART IAI Japan, and Cyber and Real Infrastructure Modeling Sub-Committee of Japan Society of Civil Engineers (JSCE).