Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals

Stephen Law, Rikuo Hasegawa, Brooks Paige, Chris Russell & Andrew Elliott

Published online: 23 May 2023.

Submit your article to this journal

Article views: 247

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS | Check for updates

Explaining holistic image regressors and classifiers in urban analytics with plausible counterfactuals

Stephen Law[a,b] (iD), Rikuo Hasegawa[a], Brooks Paige[b,c], Chris Russell[d] and Andrew Elliott[b,e]

[a]Department of Geography, University College London, London, UK; [b]The Alan Turing Institute, UK; [c]Centre for Artificial Intelligence, University College London, London, UK; [d]Oxford Internet Institute, University of Oxford, Oxford, UK; [e]School of Mathematics & Statistics, Glasgow University, Glasgow, UK

ABSTRACT

We propose a new form of plausible counterfactual explanation designed to explain the behaviour of computer vision systems used in urban analytics that make predictions based on properties across the entire image, rather than specific regions of it. We illustrate the merits of our approach by explaining computer vision models used to analyse street imagery, which are now widely used in GeoAI and urban analytics. Such explanations are important in urban analytics as researchers and practicners are increasingly reliant on it for decision making. Finally, we perform a user study that demonstrate our approach can be used by non-expert users, who might not be machine learning experts, to be more confident and to better understand the behaviour of image-based classifiers/regressors for street view analysis. Furthermore, the method can potentially be used as an engagement tool to visualise how public spaces can plausibly look like. The limited realism of the counterfactuals is a concern which we hope to improve in the future.

1. Introduction

Most explainable computer vision xCV or more generally explainable artificial intelligence xAI methods in urban analytics, explain the response of a machine learning ML classifier or regressor via localization using heatmaps.[1] Given a picture, a heatmap-based explanation may show why a classifier has labelled it as a cyclist by identifying regions of the image that the classifier maximally depends on. If these regions do in fact contain a cyclist, a scientist making use of the system can feel more comfortable that the classifier is working as intended, and not for example, exploiting correlated features in the local context, such as if an image contains a cycle lane, in deciding if it is a motorcyclist or a cyclist.

CONTACT Stephen Law ✉ stephen.law@ucl.ac.uk

Unfortunately, for many scientific tasks, we are interested in estimating attributes that do not depend on small regions of an image. For example, in urban analytics, we may be interested in estimating the greenness or scenicness of the environment from a street photo. These estimates do not depend solely on parts of the image, but on emergent properties that come from considering the entirety of the image. For a tree classifier, the explanation should not just highlight the existing trees but also all the buildings where there could be trees. As such explanability methods such as heatmaps that highlight multiple parts of the image can be difficult to observe what the methods depend on. We refer to such problems as *holistic image tasks*.

In this paper, we show how explanations can be provided for ML systems that solve these holistic image tasks, like those in urban analytics where attributes are closely linked with each other, allowing scientists, who may not be machine learning experts, to be more confident about what these systems do and how they work. To this end, we propose a new form of plausible counterfactual explanation, that visualises how images can be altered to increase or decrease the response of the ML system. To demonstrate the helpfulness of these explanations to lay users, we apply our approach for street scenes image-based classifiers/regressors in urban systems.

1.1. The importance of explaining urban analytics models

In urban systems, streets and spaces between buildings are an integral component to our livelihood, as they provide access to leisure, enable economic activities, and help to connect people. In an attempt to leverage this, much recent progress has been made using deep learning to mine street image data. This combination of data-type and method has become popular in urban analytics and GeoAI (Ibrahim *et al.* 2021, Biljecki and Ito 2021). In no small part, this growth has been driven by advances in deep learning methods, scalable computation, and the proliferation of ground level street imagery from sources such as Google and Mapillary.

Example applications include estimating demographic profiles using vehicle types (Gebru *et al.* 2017), measuring physical changes in neighbourhoods (Naik *et al.* 2017), estimating real estate values (Law *et al.* 2018, Kang *et al.* 2021), and the estimation of various perception indicators in cities (Naik *et al.* 2014, Seresinhe *et al.* 2019). Furthermore, eye-level street view, as opposed to overhead aerial view, can provide useful information for urban planning that is often costly and difficult to collect manually. To illustrate, the activeness of a street frontage as measured by the openness and the frequency of entrances and windows, can be an indicator of safety perception (Jacobs 1961, Law *et al.* 2020), which can be easily retrieved from photos of street scenes. These methods, can distill insights and help understand, map, and better monitor our urban environments, allowing us to better plan and design future neighbourhoods.

Despite the widespread use of computer vision, little research in urban analytics has been undertaken to interpret and explain the machine learning (ML) models we use (Kakogeorgiou and Karantzalos 2021). As these computer vision systems become increasingly commoditized, more scientists with little experience of deep learning or computer vision will make use of them, and thus it is important to provide

explainability tools that allows these scientists to build intuition about how these systems work so that they are not misled by their output.

This paper explores explainability in computer vision models for street images in urban analytics. In contrast to standard heatmap-based explainability approaches that give importance scores to every location in the image, we will explore the use of an explainability approach referred to as 'counterfactual explanations' (Wachter et al. 2017), which asks what would need to be altered in the street scene for the classifier to give a different result. For example, modifications can include introducing more greenery, taller buildings or bringing in more skylight onto the street scene. Such plausible urban counterfactuals are potentially more intuitive to urban planners than other existing xML(xAI) visual explanation methods, as the visualisations of urban design scenarios are often used as a public engagement tool in practice.

Methodologically, we extend on previous works (Wachter et al. 2017), by showing how counterfactual explanations (i.e. what should be altered in this image in order for it to be classified differently) can be computed in a low-dimensional latent space induced over the mid-level responses of a deep network. The low-dimensional nature of this space means that such explanations are visually more distinctive than those editing in image space which are more similar to adversarial perturbations (Elliott et al. 2021).

In this paper, we propose a counterfactual explanation pipeline in urban analytics whose resulting image provide visual explanations for street image regressors/classifiers in an urban setting. These explanations are important as stakeholders in urban planning are increasingly reliant on these models for decision making. Our research makes the following contributions:

- Introduce an explainability approach in urban analytics that produces counterfactual visual explanations for an urban image based regressor.
- Evaluate qualitatively the distinctiveness and coherency of the counterfactual visual explanations.
- Evaluate quantitatively the explanations to ensure the counterfactuals visualisations are consistent with the regressor trained from the original images.
- Compare the explanations from our urban counterfactual method with a popular baseline saliency method in a user study.

2. Related work

Several explainable AI techniques have been proposed to visually explain computer vision models such as convolutional neural networks. Most commonly used approaches are heatmap-based visualisations which highlights areas of images that are salient with respect to the classifier decision (Simonyan *et al.* 2013, Zeiler and Fergus 2014). These approaches are primarily post-hoc methods that can be further divided into perturbation-based methods e.g. (Zeiler and Fergus 2014, Ribeiro *et al.* 2016, Fong and Vedaldi 2017) and Gradient-based approaches e.g. (Simonyan *et al.* 2013, Selvaraju *et al.* 2017). The majority of the GeoAI research that uses visual explainability methods, for example in the context of remote sensing (Kakogeorgiou and Karantzalos 2021), applies heatmap-based methods such as GradCam (Selvaraju

et al. 2017) to gain insight on computer vision classifiers to improve transparency and to ensure these models are not making erroneous inference/decisions.

Although heatmap-based visualisations can highlight regions of images that are important for the classifier decision, such methods do not show how it could change (Lang *et al.* 2021). Moreover, heatmap-based methods are not necessarily understandable, and might not be able to explain more complex scenes and concepts where multiple parts are being highlighted. A recent approach that addresses some of these challenges is through the use of generative counterfactual explanations that synthesise images that are altered towards a particular classifier outcome (Goetschalckx *et al.* 2019, Härkönen *et al.* 2020, Lang *et al.* 2021). These approaches synthesize contrastive examples by searching over and editing the latent space of a generative model such as a generative adversarial network (GAN) (Goodfellow *et al.* 2014, Mirza and Osindero 2014, Karras *et al.* 2020).

2.1. Growing use of generative models in urban analytics

There is increasing application of generative models such as GANs in urban analytics as demonstrated in a recent survey from Wu *et al.* (2022). Current use includes the generation of building footprints from a street plan (Wu and Biljecki 2022), street facade restoration (Sun *et al.* 2022), satellite image enhancement (Pham and Bui 2021), satellite-to-street cross-view synthesis (Tang *et al.* 2019) and 3D reconstruction (Kelly *et al.* 2018).

There is relatively limited related research applying or editing generative models on street imagery in urban analytics. The most relevant studies are Joglekar *et al.* (2020), which applied a GAN architecture to beautify street scenes follow by a nearest neighbour search on the edited image; Sun *et al.* (2022), that applied an image-to-image translation CycleGAN (Zhu *et al.* 2017) model to renovate building facades; and Ibrahim et al. (2021), that applied a U-Net (Ronneberger *et al.* 2015) model to visualise urban design intervention from a self-curated 'before-and-after' dataset.

The approach of Joglekar *et al.* (2020) has some similarities with the method we introduce in this paper, in that it edits the latent space of a traditional GAN. Due to the noisier edited street scene, the authors proposed a nearest neighbour search to identify more plausible images. Despite this, we observed that such nearest neighbour search can yield street scenes that are perceptually too distant from the original image, as we demonstrate and discuss in section 4.2.

Although the approaches of Ibrahim et al. (2021), Sun *et al.* (2022) synthesizes more realistic examples, they are limited by the diversity and volume of the available, self-curated, ground-truth data. As the curation of such data require substantial effort, they might be unable to produce counterfactuals in a range of magnitudes for a specific attribute. We extend this line of research by introducing a generative counterfactual explanation pipeline in urban analytics for street scenes, that does not require this additional annotated data.

3. Methods and materials

The aim of the research is to explore the utility of counterfactual explanations in the urban settings. To do this we created a pipeline which allows us to both construct a

regressor of urban images for an underlying target, i.e. trees, buildings, and to also construct counterfactuals to help to explain the regressor.

3.1. Urban counterfactual architecture

The overall machine learning pipeline can be found in Figure 1. This formulation (as well as much of recent machine learning work) uses a latent representation where we map the observed data x (images of street scenes) into a relatively small learned latent space (z), which well represents the data. We can then use this latent space to both perform regression on our target of interest (e.g. presence of trees or buildings), and to reconstruct the image. A crucial decision in this work is to use this low-dimensional space to search for counterfactuals, rather than searching directly in image space which produces imperceptible changes (Elliott et al. 2021).

The processing pipeline and the models involved in the approach are shown in Figure 1, and in the remainder of the section we will introduce them one by one. In total, four models are trained or computed in this order sequentially and separately. It consists of: (1) A generative model (in red), in this case a Variational Autoencoder (VAE), consisting of an encoder that maps an input image to a lower-dimensional latent distribution, and a generator that maps from this latent space back to image space (Kingma and Welling 2013, Doersch 2016); (2) A regression model (in orange)
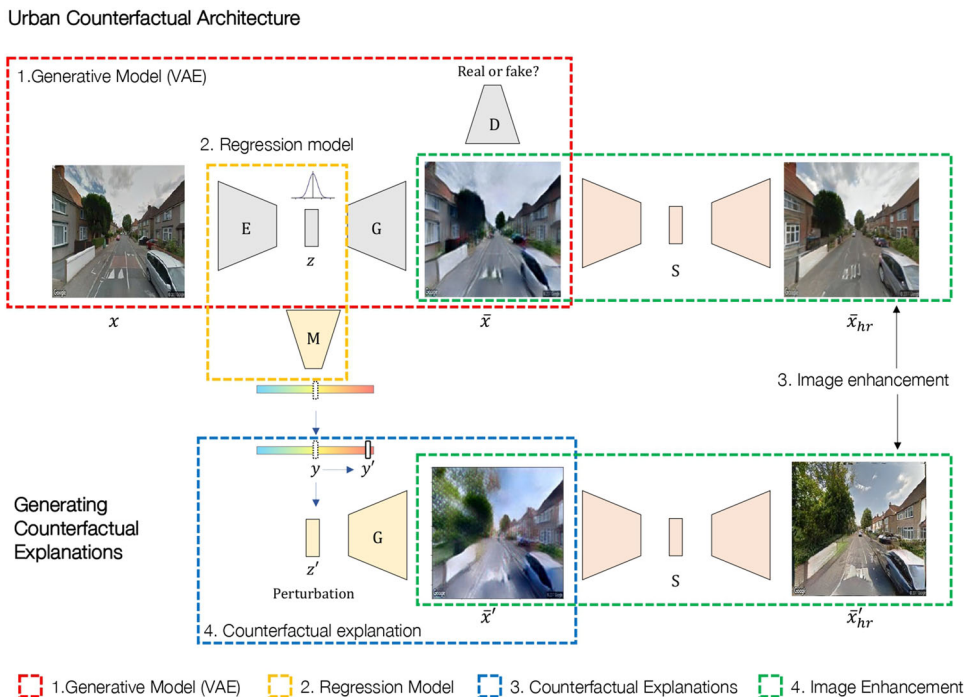

Figure 1. Pipeline for our method. It consists of (1.) A Generative model (in red) that maps an image to a latent space, (2.) a Regression model (in orange) that maps from latent space to a street scene semantic attribute, (3.) an Image enhancement model (in green) that synthesize higher quality images and (4.) a Counterfactual explanation procedure (in blue).

that maps the latent space to a target attribute; and (3) an image enhancement model (in green), that enhances the quality of the reconstruction (Choi *et al.* 2020) following recent work in image super-resolution (Umer *et al.* 2021). Finally at inference time, (4) we introduce our counterfactual explanation procedure (in blue) that allows for guided editing in the latent space for visual explanations.

The focus of the research is to demonstrate a somewhat model agnostic method for finding plausible urban counterfactuals. As a result, we have selected standard machine learning models and cite the relevant literature for each component. The generative model, regressor and image enhancement model are trained sequentially and separately. The counterfactual explanation is applied at inference time to explain the street image regressor.

3.2. Training our urban counterfactual

3.2.1. Generative model

As detailed above, our first model is a variational autoencoder (VAE) (Kingma and Welling 2013, Doersch 2016), detailed in panel 1 (red) of Figure 1. This model learns a low-dimensional latent representation that our other models can use: for regression, for reconstruction of street scenes, or (combining the two) for the creation of counterfactuals. The VAE model consists of three networks: an encoder $E(\cdot)$ which encodes the input image x to a latent space z, a corresponding generator $G(\cdot)$ that maps the latent space back to image space $\bar{x}$, and an auxiliary discriminator $D(\cdot)$ that enhances (qualitatively) the reconstruction through adversarial training. Training our adversarial VAE thus involves two stages. The first stage minimizes the VAE loss $L_{vae} = L_{rec} + L_{reg}$, where L_{rec} is the reconstruction loss and L_{reg} is the regularisation loss. The second stage follows the adversarial training procedure in LSGAN (Mao *et al.* 2017), which consists of an discriminator loss L_{dis} and a generator loss L_{gen}. The two parts of the generative model are trained sequentially using the ADAM optimizer (Kingma and Ba 2014) with a learning rate of 0.0001. Details of the architecture and the loss function can be found in the Appendix A. We have selected a VAE as it is a conventional generative model in the literature. However other generative models such as GANs (Joglekar *et al.* 2020) and diffusion models (Rombach *et al.* 2022) can be used here.

3.2.2. Classifier/regressor

The second model to train is our regressor, which is detailed in panel 2 (orange) of Figure 1. This model in essence learns the relationship between the low dimensional latent representation (from our encoder) of each of our street scenes to the underlying quantity of interest. Thus, by combining this network with the output from the trained encoder network we have a regressor that can be applied on a new image. We train a MLP regressor $M_i(\cdot)$ that maps the latent embedding z to the regression response y for attribute i, minimising the mean squared loss $L_{mse} = MSE[M_i(z), y_i]$ between the regression target and the prediction using the ADAM optimiser (Kingma and Ba 2014). The functional form for $M_i(\cdot)$ is flexible and can be a linear regressor (in our study) or a more complex non-linear one (future study). We opted for a more straightforward

linear model, as it generated plausible counterfactuals and was also more computationally efficient.

3.2.3. Image enhancement

Lastly we trained an image enhancement model, which is detailed in panel 3 (green) of Figure 1. The image-to-image enhancement/translation model uses the StarGanv2 architecture (Choi *et al.* 2020) $S(\cdot)$ to convert between multiple image domains, in this case a lower resolution image style and a higher resolution image style. The StarGanv2 contains a generator that translates an input image to a higher resolution style, a style encoder that extracts styles from an image, a mapping network that transforms a latent code into a style code, and a discriminator typical of GANs that improves image quality. We use default hyperparameters here.[2] For details of the architecture and its training details please see Choi *et al.* (2020). During inference, we can use the generator of the translation model to convert a lower resolution counterfactual $\bar{x}'$ into a higher quality one $\bar{x}'_{hr} = S(\bar{x}', hr)$. This image enhancement step ensures the reconstruction and the counterfactual is perceptually distinctive. StarGan v2 has been selected as it is a conventional architecture in the literature; other similar image2image architectures can be used here.

3.3. Generating counterfactual explanations

After training these models, we use our low dimensional embedding space to construct counterfactuals following Elliott et al. (2021), as detailed in panel 4 (blue) of Figure 1. Essentially, this procedure asks the question, 'What is the smallest change in our low dimensional representation of a street scene that would result in our regressor returning a different output?' i.e. what is the regressor reacting to in our urban imagery. Mathematically, following Elliott et al. (2021), we can find the latent position z' which gives the closest regressor prediction to T by minimising

$$\left(M_i(z') - T\right)^2 \tag{1}$$

where T is a value greater than zero and $M(\cdot)$ is a standard regressor that takes z from the latent space of the VAE. However, simply minimising Eq. (1) will not produce a low dimensional representation close to the image in question, e.g. if we ask for a lot of trees, it may return a forest rather than the requested urban scene. Thus, we add a second term which forces the latent position to be close to the latent position of the image in question:

$$\left(M_i(z') - T\right)^2 + \lambda ||z' - z||_1 \tag{2}$$

where λ balances the importance of each of the terms, thus finding a street image close to the original while changing the regressor value.

Specifically, we minimised this objective using an adaptive stochastic gradient descent optimiser (Polyak and Juditsky 1992) with a learning rate of 0.001, a $\lambda = 100$ and for 20000 iterations. The hyper-parameters for the counterfactual method have been chosen qualitatively as illustrated in Figure B1 of the Appendix B. Minimising this

objective for an appropriate value of λ and T is a good strategy for finding a plausible counterfactual.

3.4. Data augmentation

Data augmentation is a popular technique in machine learning to increase both the volume and the diversity of the training data, by applying some form of label-preserving transformation to the original data (Shorten and Khoshgoftaar 2019). Data augmentation is used here as a regulariser to reduce overfitting and to help ensure the prediction from the StarGan reconstruction is consistent with the prediction from the original data. As part of our training procedure, we test the usefulness of the augmentation by including an enhanced reconstruction of each image in the training data which results in doubling the size of the training data. Figure 2 shows the original image on the left, a reconstruction in the middle, and an enhanced reconstruction on the right synthesised for data augmentation. The enhanced reconstruction generally captures well the original image with some notable textual artefacts (e.g. road marking). The results of the data augmentation step can be seen in section 4.3.

3.5. Materials

We used the dataset of Law *et al.* (2018) consisting of street images taken from Google Streetview (Google 2017).[3] Following Law *et al.* (2018), one front-facing image was collected from the centroid of each street in the Greater London Area using the Google StreetView API. For more details, see Law *et al.* (2018). The regression target we used in this investigation is the normalised pixel counts of trees, sky and buildings extracted from a pretrained semantic segmentation model (Segnet) on street imagery which achieved a per-class average accuracy of 72% and a global accuracy of 91% for 11 classes (Badrinarayanan *et al.* 2017). The method was selected as it has been used and verified in previous urban analytic studies (Liang *et al.* 2017, Joglekar *et al.* 2020). We plan to experiment with more recent pretrained segmentation models in the future.

Figure 2. Enhanced reconstruction used for data augmentation. (left) the original street scene, (centre) street scene reconstruction, and (right) the enhanced reconstruction used for data augmentation. Contains Google StreetView data © Google © 2017.

4. Evaluation and results

In order to evaluate our explainability method, we conducted both qualitative evaluations demonstrating the results of our approach, and quantitative evaluations including a user study to empirically verify our method.

4.1. Qualitative evaluation

We first visually inspect the counterfactuals produced by our approach. Figure 3 shows the original image (left-most column) and a generated counterfactual for perturbing each of the different regression attributes in turn – showing an increase in the number of tree pixels, building pixels and sky pixels, respectively. For each selected street image, we see that the target regression attribute is visually distinctive, and that the
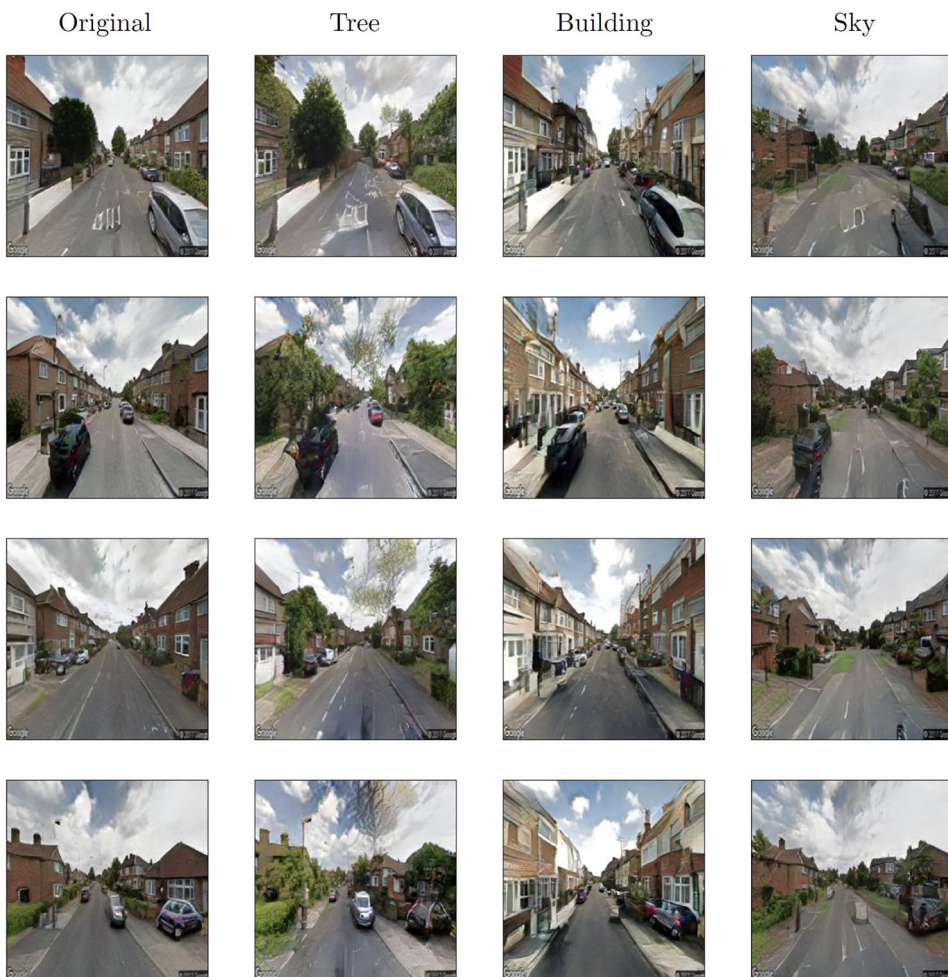

Figure 3. Figure showing the urban counterfactual for each regressor target (Tree, Building, Sky) for four example images. Contains Google StreetView data [copyright] Google [copyright] 2017.

$\hat{y}_{tree} = 0.064$ $\hat{y}_{tree} = 0.115$ $\hat{y}_{tree} = 0.197$ $\hat{y}_{tree} = 0.231$ $\hat{y}_{tree} = 0.263$ $\hat{y}_{tree} = 0.259$

$\hat{y}_{bldg} = 0.154$ $\hat{y}_{bldg} = 0.235$ $\hat{y}_{bldg} = 0.300$ $\hat{y}_{bldg} = 0.353$ $\hat{y}_{bldg} = 0.409$ $\hat{y}_{bldg} = 0.448$

$\hat{y}_{sky} = 0.202$ $\hat{y}_{sky} = 0.238$ $\hat{y}_{sky} = 0.272$ $\hat{y}_{sky} = 0.293$ $\hat{y}_{sky} = 0.301$ $\hat{y}_{sky} = 0.306$

Figure 4. From left to right, showing how the urban counterfactual shifts towards its regression attributes for varying levels of $T \in \{0, 1, 2, 3, 4, 5\}$. The regressor prediction ($\hat{y}$) presented are computed on the full counterfactual reconstruction (i.e. the image shown). Contains Google StreetView data [copyright] Google [copyright] 2017.

perturbation are localised in expected regions of the image. For example, tree pixels are added on the sides of streets, building pixels are added above the existing buildings, and sky pixels are used to remove existing buildings. However, there are still significant artefacts in the perturbed image; for example, a lack of windows on the transformed buildings, unrealistic road markings and shadows of trees emerging from the sky.

To study how the perturbation changes for different levels of regression targets, we computed a set of explanations where we set $\lambda = 100$ and vary $T \in \{0, 1, 2, 3, 4, 5\}$ in Eq. (2). Figure 4 shows how a street scene gradually shift towards each of its regression attributes as predicted from the StarGan reconstruction $\hat{y}$ (i.e. Tree, Sky, Building) as T increases.

A prominent example is the building counterfactual where the trees on the left-side are slowly being removed before a floor is added on the right-side. These result also generally show perturbing with a larger magnitude is necessary for the model to make perceptually meaningful counterfactuals.

However, its realism also reduces significantly at higher levels of T suggesting these generated images might lie outside the data manifold. This is most apparent for the tree counterfactual where the road markings become distorted when $\hat{y}_{tree} > 0.231$. We also notice that multiple attributes are being edited concurrently. The most notable example is the sky counterfactuals where the vehicle begins to disappear when $\hat{y}_{sky} > 0.301$. These results suggest our perturbations are making changes to the images holistically, but it also do not fully disentangle each semantic attribute when multiple attributes change concurrently. In such cases, street scenes with more sky pixels are more likely to have lower building density and consequentially fewer vehicles.

4.2. Nearest neighbours comparison

Next we compared our approach to a previous one, Joglekar *et al.* (2020), by visualising the counterfactuals in our manner as oppose to looking for street scenes in the dataset that are closest to it. Figure 5 shows an original street scene, a counterfactual generated from our approach with an increase number of tree pixels, and the two nearest neighbours in the latent space z of the counterfactuals. Despite showing some similarities between the counterfactuals and the nearest neighbours, such as the vegetation in both sets of images, there are notable differences between the two street scenes in terms of architecture style and the vehicles on the street. These results are similar to those observed in Joglekar *et al.* (2020) where the high dimensionality of images means that finding an exact counterfactual street scene is difficult. Despite some notable artefacts such as changes on the road marking and the sky, the generated street scene is visually closer to the original street scene than its nearest neighbours.

4.3. Quantitative evaluation

To evaluate whether the counterfactuals produced by the model pipeline are consistent with the prediction of the image regressor, we trained three sets of models. The first set is between the observed and the predicted from the original image, with and without data augmentation. The second set is between the observed and the predicted from the StarGan reconstruction, with and without data augmentation. The third set is between the predicted from the original image and the predicted from the StarGan reconstruction, with and without data augmentation.

For all experiments, we divided the dataset ($N = 20{,}000$) into a train- and test-set (80 : 20) where we train a linear regressor $M_i(\cdot)$ for each attribute i that minimises the mean squared error on the training set, using the ADAM optimiser (Kingma and Ba 2014). We tested with learning rates 0.001 and 0.0001 and report the test set R^2, *MSE*
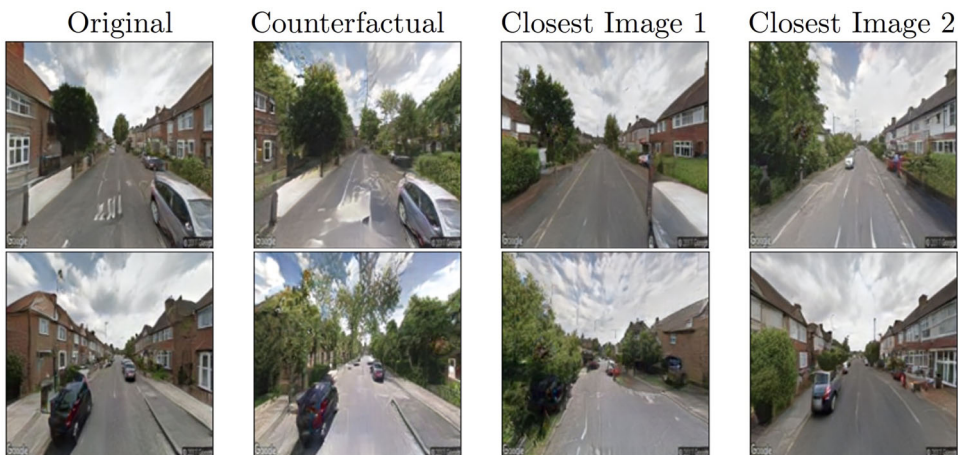

Figure 5. Nearest neighbours for the counterfactuals. Contains Google StreetView data [copyright] Google [copyright] 2017.

and *MAE* between the ground truth, the predicted from the original image and the predicted from the StarGan reconstruction.[4]

Table 1 shows the goodness of fit between the ground-truth (true), predicted from the original image (pred) and the predicted from the StarGan reconstruction (pred SG), with and without data augmentation (trained with $lr = 0.0001$). These results display a positive fit between the ground truth and the predicted from the original image. The fit is stronger between the ground truth and the predicted from the original image as opposed to the ground truth and the predicted from the StarGan reconstruction. The gap decreases when using the data augmentation pipeline suggesting the response predicted from the counterfactual is more consistent when using the augmented regressor. However, there is also a reduction in fit between the ground-truth and predicted from the original image when using the data augmentation pipeline. Results for model trained with $lr = 0.001$ (in table B1 of the Appendix B) shows similar findings but with a larger gap between the R^2 of the sky and tree attributes. The results show a reasonably strong fit and consistency between the observed and the predicted from the image and the reconstruction. Future research is necessary to improve the overall fit of the image regressor.

4.4. User study

In order to evaluate whether our approach can generate distinctive and plausible visual explanations, we performed four user experiments using Amazon Mechanical Turk (a popular crowd survey tool), with a protocol adapted from Lang *et al.* (2021) and compared our method with GradCam (Selvaraju *et al.* 2017), a popular heatmap-based method for explaining image-based holistic image classifiers/regressors. For our approach, we computed a set of explanations where we set $\lambda = 100$, $T_{tree} = 1.5$, $T_{building} = 3$ and $T_{sky} = 5$ (parameters set based on qualitative results) in Eq. (2). For the baseline method, we used the Captum (Kokhlikyan *et al.* 2020) interpretability library to produce GradCam heatmaps. We utilised default settings, visualising the last *ReLU* layer before the linear regressor.[5]

4.4.1. Experiment 1: Street scenes change survey
For the first user experiment, each participant is shown four pairs of images, corresponding to modification of one of the three attributes (tree, building, sky) for a given

Table 1. Statistical analysis results for the regression models trained with $lr = 0.0001$.

		true vs pred		true vs pred SG		pred vs pred SG	
		Orig.	Aug.	Orig.	Aug.	Orig.	Aug.
Tree	R^2	0.717	0.721	0.167	0.669	0.127	0.780
	MSE	0.0008	0.0008	0.0025	0.0010	0.0022	0.0005
	MAE	0.0222	0.0220	0.0400	0.0240	0.0388	0.0178
Building	R^2	0.753	0.737	0.188	0.696	0.219	0.833
	MSE	0.0021	0.0021	0.0069	0.0026	0.0056	0.0012
	MAE	0.0357	0.0372	0.0686	0.0398	0.0648	0.0269
Sky	R^2	0.806	0.779	0.447	0.734	0.577	0.861
	MSE	0.0007	0.0008	0.0019	0.0009	0.0013	0.0005
	MAE	0.0200	0.0215	0.0357	0.0236	0.0303	0.0171

We report the test set R^2, *MSE* and *MAE* between the ground truth (true), the predicted from the original image (pred) and the predicted from the StarGan reconstruction (pred SG).

image. The two transforms on the left both modify the same unknown attribute i, serving as a baseline. On the right, one transform is shown which modifies i in a similar magnitude, alongside another transform which instead modifies attribute j. The user is then asked to identify which of the two transforms on the right (option A or B) matches the two baseline transforms on the left. A correct response is when the participant selects a transformation that corresponds to modifying the same attribute i. An example of this experiment can be found in the top panel of Figure 6.

4.4.2. Experiment 2: Street scenes heatmap survey

For the second user experiment, each participant is similarly shown four pairs of images, but this time rather than showing a transformation, it instead shows a GradCam heatmap that highlights relevant regions for one of the three attributes (tree, building, sky) in a given image. As in experiment 1, the left two heatmaps both correspond to a particular attribute i, whereas on the right, one of the heatmap is produced from the same attribute i while the other is produced from a different attribute j. The user is then asked to identify the heatmap on the right (option A or B) that corresponds to the same attribute i as highlighted on the left. This experiment is used as a comparison for Experiment 1. An example of this experiment can be found in in the bottom panel of Figure 6.
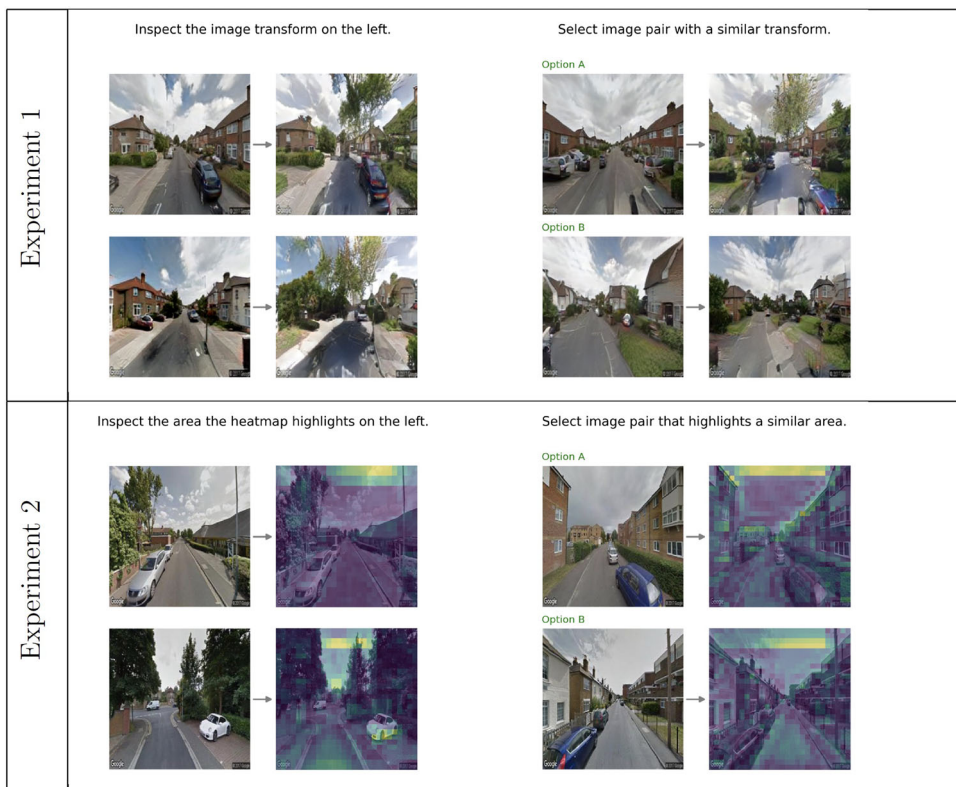

Figure 6. **Top:** experiment 1 **Bottom:** experiment 2 Contains Google StreetView data [copyright] Google [copyright] 2017.

4.4.3. Experiment 3: Street scenes change description

For the third user experiment, each participant is shown two rows of images. The top row is a set of four randomly selected street imagery and the bottom row are counterfactuals produced from one of the attribute i (tree, building, sky). The user is then asked to describe in 1–4 words the single most prominent attribute they see increasing for all the images. A correct response is when the participant describes the transformation that corresponds to the attribute i being transformed. An example of this experiment can be found in in the top panel of Figure 7.

4.4.4. Experiment 4: Street scenes heatmap description

For the fourth user experiment, each participant is shown two rows of images. The top row is a set of four randomly selected street imagery and the bottom row are heatmaps corresponding to one of the attributes i (tree, building, sky). The user is then asked to describe in 1–4 words the single most prominent attribute that is being highlighted for all the heatmaps. An example of this experiment can be found in the bottom panel of Figure 7.

For the user study experiments, we used the augmented regressor with $lr = 0.001$ (see Table B1) and constructed 120 sets of reconstructions, counterfactual explanations and GradCam heatmaps for each attribute i. We then randomly selected images (with replacement) and made 40 tasks for each experiment and ran four separate experiments for seven days where each worker would get the same sets of questions. Following an initial test, we hired experienced AMT workers/participants (i.e. those ranked as 'Mechanical Turk Masters') for the user study to ensure higher quality of the submissions. Manual data cleaning was necessary for experiment 3 and experiment 4 to remove erroneous textual descriptions by a small number of users (e.g. using a mathematical transform as a description). A total of 50 were removed for experiment 3, and 134 were removed for experiment 4. In the end we coded and approved 704 responses for experiment 1, 682 responses for experiment 2, 840 responses for experiment 3 and 780 responses for experiment 4.[6]

Table 2 shows a summary table (accuracy and standard error in brackets) for all four experiments. The result shows that users with access to our explanations have higher accuracy than GradCam for both sets of surveys indicating that our approach gives more distinctive explanations for these street image classifiers.

Table 3 shows the summary table for both the street scene change survey using our method on the left and the street scene heatmap survey using the baseline method on the right. Users with access to our explanations (left) have greater accuracy for the tree and building counterfactuals and a slightly lower accuracy for the sky counterfactuals suggesting the challenges for interpreting the background attribute (sky) relative to the foreground attribute. The baseline gradcam method results in a lower user accuracy than our approach for all three attributes; with multiple parts of the image highlighted (see Figures 6 and 7). These results suggest our approach provides more distinctive and noticeable explanations for holistic image classifiers which depend on the entirety of the image.

Table 4 shows the summary table for street scene change description task using our approach on the left and street scene heatmap description using the baseline
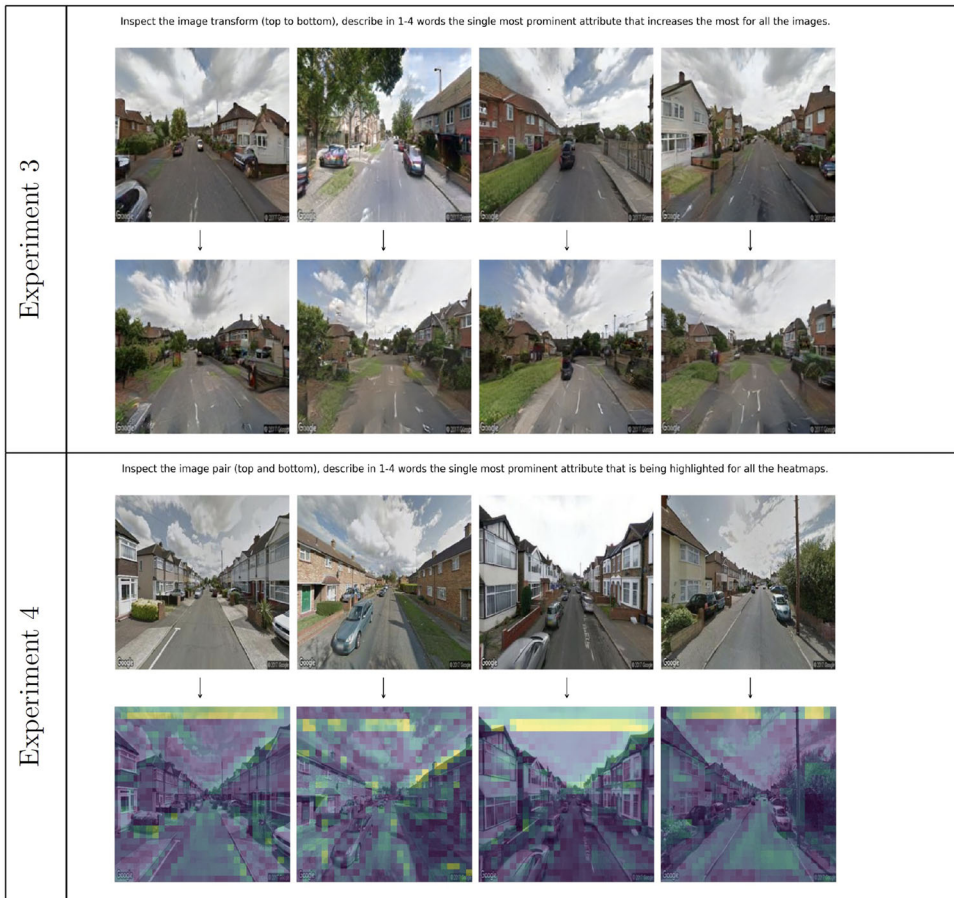
Figure 7. Top: experiment 3 **bottom:** experiment 4 Contains Google StreetView data [copyright] Google [copyright] 2017.

Table 2. User study results summary.

Experiment	Accuracy (± S.E.)	N
01 Urban Counterfactual Survey (Ours)	0.695 (± 0.017)	704
02 Urban Heatmap Survey (GradCam)	0.543 (± 0.019)	682
03 Urban Counterfactual Desc (Ours)	0.601 (± 0.017)	840
04 Urban Heatmap Desc (GradCam)	0.513 (± 0.018)	780

Table 3. Results for urban change and heatmap survey.

Urban explanation	Counterfactual (Ours) (Exp. 1)		GradCam (Exp. 2)	
Survey	Accuracy (± S.E.)	N	Accuracy (± S.E.)	N
tree	0.690 (± 0.029)	252	0.498 (± 0.031)	257
building	0.787 (± 0.028)	211	0.568 (± 0.033)	220
sky	0.618 (± 0.031)	241	0.571 (± 0.035)	205

saliency method on the right. Our method (left) shows a higher accuracy for the tree and building attribute and a lower accuracy for the sky attribute. The main reason being that keywords such as increase of 'sky' or 'clouds' are not used when describing

Table 4. Results for urban change and heatmap description.

Urban explanation	Counterfactual (Ours) (Exp. 3)		GradCam (Exp. 4)	
Description	Accuracy (± S.E.)	N	Accuracy (± S.E.)	N
tree	0.824 (± 0.022)	295	0.598 (± 0.029)	286
building	0.723 (± 0.028)	253	0.280 (± 0.026)	293
sky	0.271 (± 0.026)	292	0.731 (± 0.031)	201

a street scene with more sky pixels. Instead, the participants describe the scenes as having 'less buildings' or 'less trees' or 'emptier streets'. The baseline method (right) shows opposite results where the sky attribute achieved a higher accuracy relative to the building and tree attribute. In contrast, keywords such as 'sky' or 'clouds' are often used when describing a street scene where the sky pixels are highlighted. These results potentially suggest that heatmap localisation methods can be more perceptible when describing a background attribute than our approach.

To briefly summarise the user study, these results generally suggest our approach can help better understand the use of machine learning model for describing holistic computer vision regressors for street scene analysis. Qualitatively, the direction and the magnitude of the change is less detectable with the heatmap approaches where multiple parts of the image is being highlighted. Whereas for our approach, the modified attribute is visualised, in that our explanation for a tree classifier will show more trees in the street scene. However, a couple of limitations are observed with our approach for the user experiment. One is that changes with background attributes can be perceptually less evident when describing the changes of a street scene with our approach (e.g. more sky vs less buildings) and in instances whose attributes have more extreme values, perturbing in the same direction as these attribute might also be less noticeable (e.g. increasing trees in an already forested street).

5. Discussion and concluding remarks

Despite the rising popularity of utilising ML in Urban Analytics (Biljecki and Ito 2021), there is presently limited research focusing on explaining machine learning models in this domain. To address this we propose a novel pipeline to explain holistic image regressors/classifiers in urban analytics, by synthesising counterfactuals, where the explanations are computed over the lower dimensional latent space of a deep neural network as oppose to image space (Elliott et al. 2021). To illustrate and verify our novel pipeline, we applied our methodology to a set of Google StreetViews (GSV) images of London for multiple regression targets. We validated our approach, both qualitatively through visual comparisons and quantitatively through a user study comparing our approach to an existing baseline heatmap method. Through this study, we were able to demonstrate that our explanation is distinctive and can be better understood by non-experts.

The development of explanations/explainable models are very important in urban systems, as urban planners are increasingly reliant on these types of models for decision making. As an illustration, given a model that uses street images to predict house prices (Law et al. 2018, Kang et al. 2021), one can visualise the changes in the street scenes due to an increase in price, thereby revealing the features in the scene that

correspond with the regressor, which could correspond to architectural styles or street greenery. Further, these visual counterfactuals are potentially useful in urban planning beyond aiding in the understanding of computer vision systems in street view analysis. They can also serve as a visual communication tool to help stakeholders imagine how future public spaces can plausibly look like, e.g. what if we add more greenery, denser buildings or even a cycle-lane on a particular street.

Methodologically, one benefit of our approach is its flexibility, in that a new regression target (e.g. house price or cycle-lanes or semantic classes such as pedestrians and vehicles from the Cityscape dataset)[7] can be easily fitted without needing to retrain the entire model from scratch. A further advantage of our approach is the introduction of a generic pipeline for generating urban counterfactuals, which allows us to replace the various components (see Figure 1) as needed. However, one consequence of this is we did not test each model component exhaustively resulting in artefacts and limited realism for the generated street scenes. Thus, with the growing popularity of generative models for realistic image synthesis (Rombach *et al.* 2022), our pipeline can be adapted for use with these newer methods for future research.

In this work, we focused on the use of Google StreetViews (GSV), a widely used data source in urban analytics (Biljecki and Ito 2021, Ibrahim *et al.* 2021). GSV offers several advantages for our work, namely, data consistency and coverage which potentially can help generate less noisy and geographically more diverse street scenes when compared to open data sources such as Mapillary. However, the use of such data sources also has its own drawbacks, i.e. the inability to share data and pretrained models which leads to a lack of reproducibility. In the future, we plan to explore open data sources using our pipeline.

Finally, it is important to acknowledge that these methods pose increasing ethical risks in geography, such as the possibility of satellite imagery being manipulated for malicious purposes (Zhao *et al.* 2021). Therefore, it is essential to conduct research and develop techniques that can detect 'deep fake' street scene to prevent misuse in the future.

Notes

1. In this work, we use image-based classifier and regressor interchangeably.
2. We trained for 100,000 iterations using the ADAM optimizer (Kingma and Ba 2014) with default hyper-parameters $\lambda_{sty} = 1$, $\lambda_{ds} = 1$ and $\lambda_{cyc} = 1$ following (Choi *et al.* 2020).
3. ©2017 Google Inc. Google and the Google logo are registered trademarks of Google Inc.
4. We tested other learning rates (0.1 and 0.01) but these models did not converge.
5. See https://github.com/pytorch/captum for more information.
6. The user study has been approved by the university departmental internal ethics review where further formal reviews were not instructed. All results were suitably anonymised where researchers on this project have no access to personal identifiable information that is held by Amazon Mechanical Turk.
7. https://www.cityscapes-dataset.com/

Acknowledgement

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Notes on contributors

Stephen Law is a Lecturer for the MSc Social and Geographic Data Science programme at University College London Department of Geography and a Fellow at the Alan Turing Institutes. He is interested in applying data driven and explainable machine learning methods in urban design, social science and geography.

Rikuo Hasegawa is a software engineer who completed a BSc Physics in UCL. He is interested in all things related to virtual reality, real-time computer graphics and type theory.

Brooks Paige is an associate professor in machine learning at the University College London AI Centre and a fellow at the Alan Turing Institute. He is interested in probablistic programming and developing interpretable machine learning models which complement human expertise, rather than attempt to replace it.

Chris Russell is a scientist at the Oxford Internet Institute as well as a senior applied scientist in Amazon Web Service. He has a particular interests on computer vision, algorithmic fairness and explainable AI.

Andrew Elliott is a lecturer in statistics at the University of Glasgow and a Fellow at the Alan Turing Institute. He is interested in the development of scalable network-based algorithms and explainable machine learning methods to solve various real world problems.

ORCID

Stephen Law ⓘ http://orcid.org/0000-0003-3184-572X

Data and codes availability statement

The code that support the findings of this study is available here https://github.com/booboo18/PlausibleUrbanCounterfactual. The street image data are not publicly available. As a result, mocked data have been provided. To run the urban counterfactuals, a trained generative model and regressive is required.

References

Badrinarayanan, V., Kendall, A., and Cipolla, R., 2017. Segnet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39 (12), 2481–2495.

Biljecki, F., and Ito, K., 2021. Street view imagery in urban analytics and gis: a review. *Landscape and Urban Planning*, 215, 104217.

Choi, Y., *et al.*, 2020. Stargan v2: diverse image synthesis for multiple domains. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Doersch, C., 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*.

Elliott, A., Law, S., and Russell, C., 2021. Explaining classifiers using adversarial perturbations on the perceptual ball. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10693–10702.

Fong, R.C., and Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation. In: *Proceedings of the IEEE international conference on computer vision*, 3429–3437.

Gebru, T., et al., 2017. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences of the United States of America*, 114 (50), 13108–13113.

Goetschalckx, L., et al., 2019. Ganalyze: toward visual definitions of cognitive image properties. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5744–5753.

Goodfellow, I., et al., 2014. Generative adversarial nets. *In*: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger, eds. *Advances in neural information processing systems 27*. New York: Curran Associates, Inc., 2672–2680. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Google. 2017., Google StreetViews extracted in 2017 from https://www.maps.google.com/.

Härkönen, E., et al., 2020. Ganspace: discovering interpretable Gan controls. *Advances in Neural Information Processing Systems*, 33, 9841–9850.

Ibrahim, M.R., Haworth, J., and Christie, N., 2021. Re-designing cities with conditional adversarial networks. *arXiv preprint arXiv:2104.04013*.

Ibrahim, M.R., Haworth, J., and Cheng, T., 2020. Understanding cities with machine eyes: a review of deep computer vision in urban analytics. *Cities*, 96, 102481.

Jacobs, J., 1961. *The death and life of great American cities*. New York: Random House Inc.

Joglekar, S., et al., 2020. Facelift: a transparent deep learning framework to beautify urban scenes. *Royal Society Open Science*, 7 (1), 190987.

Johnson, J., Alahi, A., and Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: *European conference on computer vision*. Berlin: Springer, 694–711.

Kakogeorgiou, I., and Karantzalos, K., 2021. Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 103, 102520.

Kang, Y., et al., 2021. Understanding house price appreciation using multi-source big geo-data and machine learning. *Land Use Policy*, 111, 104919.

Karras, T., et al., 2020. Analyzing and improving the image quality of stylegan. *In*: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,*. 8110–8119.

Kelly, T., et al., 2018. Frankengan: guided detail synthesis for building mass-models using style-synchonized gans. *arXiv preprint arXiv:1806.07179*.

Kingma, D.P., and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv: 1412.6980*.

Kingma, D.P., and Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv: 1312.6114*.

Kokhlikyan, N., et al., 2020. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*.

Lang, O., et al., 2021. Explaining in style: training a gan to explain a classifier in stylespace. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 693–702.

Law, S., et al., 2020. Street-frontage-net: urban image classification using deep convolutional neural networks. *International Journal of Geographical Information Science*, 34 (4), 681–707.

Law, S., Paige, B., and Russell, C., 2018. Take a look around: using street view and satellite images to estimate house prices. Available from: https://arxiv.org/abs/1807.07155.

Liang, J., et al., 2017. Automatic sky view factor estimation from street view photographs – a big data approach. *Remote Sensing*, 9 (5), 411.

Mao, X., et al., 2017. Least squares generative adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*, 2794–2802.

Mirza, M., and Osindero, S., 2014. Conditional generative adversarial nets. *CoRR*, abs/1411.1784. Available from: http://arxiv.org/abs/1411.1784.

Naik, N., *et al.*, 2014. Streetscore-predicting the perceived safety of one million streetscapes. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 779–785.

Naik, N., *et al.*, 2017. Computer vision uncovers predictors of physical urban change. *Proceedings of the National Academy of Sciences of the United States of America*, 114 (29), 7571–7576.

Pham, V.D., and Bui, Q.T., 2021. Spatial resolution enhancement method for landsat imagery using a generative adversarial network. *Remote Sensing Letters*, 12 (7), 654–665.

Polyak, B.T., and Juditsky, A.B., 1992. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30 (4), 838–855.

Ribeiro, M.T., Singh, S., and Guestrin, C., 2016. "Why should I trust you?" explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

Rombach, R., *et al.*, 2022. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O., Fischer, P., and Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. Berlin: Springer, 234–241.

Selvaraju, R.R., *et al.*, 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*, 618–626.

Seresinhe, C.I., *et al.*, 2019. Happiness is greater in more scenic locations. *Scientific Reports*, 9 (1), 1–11.

Shorten, C., and Khoshgoftaar, T.M., 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6 (1), 1–48.

Simonyan, K., Vedaldi, A., and Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K., and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sun, C., Zhou, Y., and Han, Y., 2022. Automatic generation of architecture facade for historical urban renovation using generative adversarial network. *Building and Environment*, 212, 108781.

Tang, H., *et al.*, 2019. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2417–2426.

Umer, R.M., Munir, A., and Micheloni, C., 2021. A deep residual star generative adversarial network for multi-domain image super-resolution. In: *2021 6th International Conference on Smart and Sustainable Technologies (SpliTech)*. IEEE, 01–05.

Wachter, S., Mittelstadt, B., and Russell, C., 2017. Counterfactual explanations without opening the black box: automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31, 841.

Wu, A.N., and Biljecki, F., 2022. Ganmapper: geographical data translation. *International Journal of Geographical Information Science*, 36 (7), 1394–1422.

Wu, A.N., Stouffs, R., and Biljecki, F., 2022. Generative adversarial networks in the built environment: a comprehensive review of the application of gans across data types and scales. *Building and Environment*, 223, 109477.

Zeiler, M.D., and Fergus, R., 2014. Visualizing and understanding convolutional networks. In: *European conference on computer vision*. Berlin: Springer, 818–833.

Zhao, B., *et al.*, 2021. Deep fake geography? When geospatial data encounter artificial intelligence. *Cartography and Geographic Information Science*, 48 (4), 338–352.

Zhu, J.Y., *et al.*, 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Appendix A. Model architecture and training details

The adversarial VAE model comprises of three components: an encoder $E(\cdot)$ that encodes an image x into a latent space z, a generator $G(\cdot)$ that maps from latent space back to image space $\bar{x}$, and an auxiliary discriminator $D(\cdot)$ that fine-tunes the reconstruction by distinguishing whether an image is real x or not real $\bar{x}$. These models are sequentially trained over two stages. The first stage minimizes the VAE loss $L_{vae} = L_{rec} + L_{reg}$, where L_{rec} is the reconstruction loss and L_{reg} is the regularisation loss. L_{rec} is calculated as the perceptual distance between the image features of the input image x and the reconstructed image $\bar{x} = G(E(x))$ Johnson *et al.* (2016). Formally, $L_{rec} = \sum_{l \in \mathcal{L}} ||C^{(l)}(x) - C^{(l)}(\bar{x})||_2^2$ where $C^{(l)}(x)$ is the classifier response of the lth layer of a pretrained *VGG*16 (Simonyan and Zisserman 2014). L_{reg} is calculated as the KL divergence D_{KL} between the real and the encoder as described in Kingma and Welling (2013). Formally, $D_{KL} = -0.5 \sum (1 + \log(\sigma^2) - (\mu)^2 - (\sigma^2))$ where μ and σ^2 are the mean and the variance of the latent space from the Encoder $E(x)$. The second stage follows the adversarial training procedure in LSGAN (Mao *et al.* 2017), which consists of a discriminator loss L_{dis} and a generator loss L_{gen}. More formally, $L_{dis} = 0.5(\frac{1}{n}\sum((D(x) - 1)^2) + \frac{1}{n}\sum^{n}(D(\bar{x})^2)$ and $L_{gen} = 0.5(\frac{1}{n}\sum((D(\bar{x}) - 1)^2))$, where x is the input image, $\bar{x}$ is the reconstructed image and D is the discriminator.

Table A1. Encoder.

Layer	Resample	Norm	Output
In	–	–	$224 \times 224 \times 3$
Con	–	Ins	$224 \times 224 \times 64$
Con	Max	Ins	$112 \times 112 \times 64$
Con	–	Ins	$112 \times 112 \times 64$
Con	Max	Ins	$56 \times 56 \times 64$
Con	–	Ins	$56 \times 56 \times 64$
Con	Max	Ins	$28 \times 28 \times 64$
Con	–	Ins	$28 \times 28 \times 64$
Con	Max	Ins	$14 \times 14 \times 64$
Lin	–	–	1568×2
Out	–	–	1568

Table A2. Generator.

Layer	Resample	Norm	Output
In	–	–	1568
Lin	–	–	12544
Con	Up	Ins	$28 \times 28 \times 64$
Con	–	Ins	$28 \times 28 \times 64$
Con	Up	Ins	$56 \times 56 \times 64$
Con	–	Ins	$56 \times 56 \times 64$
Con	Up	Ins	$112 \times 112 \times 64$
Con	–	Ins	$112 \times 112 \times 64$
Con	Up	Ins	$224 \times 224 \times 64$
Out	–	–	$224 \times 224 \times 3$

Table A3. Discriminator.

Layer	Resample	Norm	Output
In	–	–	$224 \times 224 \times 3$
Conv	Max	Ins	$112 \times 112 \times 28$
Conv	Max	Ins	$56 \times 56 \times 36$
Conv	Max	Ins	$28 \times 28 \times 48$
Conv	Max	Ins	$14 \times 14 \times 64$
Conv	Max	Ins	$7 \times 7 \times 64$
Out	–	–	1

Appendix B. Extra fit results and visualisations

Table B1. Statistical analysis results for model trained with $lr = 0.001$.

		true vs pred		true vs pred SG		pred vs pred SG	
		Orig.	Aug.	Orig.	Aug.	Orig.	Aug.
Tree	R^2	0.716	0.529	0.288	0.554	−0.353	0.781
	MSE	0.0008	0.0014	0.0021	0.0013	0.0026	0.0004
	MAE	0.0218	0.0292	0.0380	0.0281	0.0458	0.0156
Building	R^2	0.652	0.751	0.078	0.705	0.516	0.843
	MSE	0.0030	0.0021	0.0079	0.0025	0.0028	0.0009
	MAE	0.0433	0.0364	0.0746	0.0395	0.0430	0.0232
Sky	R^2	0.805	0.564	0.462	0.443	0.675	0.900
	MSE	0.0007	0.0015	0.0019	0.0019	0.0008	0.0003
	MAE	0.0202	0.0330	0.0356	0.0376	0.0238	0.0125

Figure B1. Urban counterfactuals varying λ and no. of iteration.