Perceptive Visual Urban Analytics is Not (Yet) Suitable for Municipalities

Tim Alpherts
t.o.l.alpherts@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Sennay Ghebreab
s.ghebreab@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Yen-Chia Hsu
y.c.hsu@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

Nanne Van Noord
n.j.e.vannoord@uva.nl
University of Amsterdam
Amsterdam, The Netherlands

ABSTRACT

The use of Computer Vision, through a Perceptive Visual Urban Analytics (VUA) paradigm, has been proposed as a way for municipalities to more easily monitor their cities. However, prior studies fall short of actually investigating whether Perceptive VUA is ready for municipal use. In this paper we take a critical look at this paradigm by comparing key methods and evaluating them on usability and trustworthiness with municipal experts as well as Responsible AI and Computer Vision researchers. Based on on this evaluation we find that Perceptive VUA is not (yet) ready for municipal use as they do not incorporate domain knowledge and overly rely on spurious correlations. We conclude by providing recommendations for how to progress Perceptive VUA such that it may actually contribute to improving the liveability and quality of urban environments.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**; • **Applied computing**;

KEYWORDS

Explainability, Computer Vision, Trustworthiness

ACM Reference Format:
Tim Alpherts, Sennay Ghebreab, Yen-Chia Hsu, and Nanne Van Noord. 2024. Perceptive Visual Urban Analytics is Not (Yet) Suitable for Municipalities. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3630106.3658976

1 INTRODUCTION

Studies on urban perception have been around since the 1960s [21], attempting to shed light on the way cities are perceived by citizens. Motivated by potential health outcomes, municipalities have also been focused on improving neighbourhoods to benefit their citizens. More recently, this field of research has caught the eye of computer vision researchers [2, 7, 18, 27, 28, 34, 38]. Perceptive VUA (Visual Urban Analytics), categorised as a subfield of VUA by [15], has attempted to evaluate whether large scale computer vision approaches can effectively predict labels from datasets of urban imagery. These approaches typically use a dataset of Google Street View images in combination with either socio-economic labels collected through civil databases such as housing prices [18] and mean income [38], or subjective labels collected through crowdsourcing such as perceived safety [27] and scenicness [34]. These approaches are motivated by the labour intensive nature of their traditional counterpart; surveys take up a lot of time and resources, and thus it is hard for municipalities to set them up at a large spatial scale. Perceptive VUA researchers pose that with a VUA approach, the perceptive VUA paradigm can learn large spatio-temporal patterns to evaluate policies [37, 38, 43] or assist urban planners in surveying the city [7, 27, 34, 42]. While there is an increasing demand from municipalities, and the public sector in general, for adopting such techniques [4, 24, 25, 30, 41] there is little research on evaluating whether the current paradigm is yet suitable for practical use within a municipal context.

As such, in this paper we evaluate whether the current paradigm of perceptive VUA for predicting socio-economic variables using street view imagery is suitable yet for incorporating into the work of municipalities and policymakers. A key requirement for integrating perceptive VUA into the municipal workflow is to combine it with explainability approaches, as a solely predictive model would be of little use for municipalities. As such we combine the perceptive VUA paradigm with explainability techniques. We compare how well the perceptive VUA paradigm, combined with different explainability approaches, helps us understand the relation between visual elements of a panoramic image and a socio-economic variable in the city of Amsterdam. To this end we compare three approaches that in our view cover the current field of Perceptive VUA: a high impact traditional approach using clustered visual elements [2, 6], an end-to-end approach as done by [38] in combination with a post-hoc explainability technique, and an end-to-end approach using the same philosophy as the previous but with the inherently explainable ProtoPNet. Additionally, we visually inspect the explanations to evaluate whether they show human interpretable visual elements and whether the methods are trustworthy. Moreover, we evaluate the methods through an expert user-study with Computer Vision and Responsible AI researchers, and employees from the

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658976

municipality that work in AI or contribute to policy-making in tech innovation. Our contributions are as follows:

- An evaluation of the current paradigm of Perceptive VUA through a comparison of existing methods on interpretability, trustworthiness and practicality for use within a municipality.
- An expert evaluation of the Perceptive VUA methods with employees from the Municipality of Amsterdam and Rotterdam, Computer Vision Researchers, and Responsible AI Researchers.
- Recommendations on usability and trustworthiness of models for obtaining insights about urban visual relationships for policy-making and future research directions for Perceptive VUA.

2 RELATED WORK

2.1 Visual Urban Analytics

The field of VUA can roughly be subdivided into two categories: Methods used as a tool to detect visual elements for which the relevancy is based on domain knowledge, and methods that identify the relevant visual elements themselves. An example of the first category could be pothole detection [22]. Road maintenance is costly and a model trained to detect potholes can be built by training an object detection model on a dataset of roads with bounding boxes around potholes. In this domain, knowledge is applied to force a model to focus on a specific part of the image (i.e., potholes). Similar methods can be applied for detecting trash [39], disparities in police deployment [10], or the number of trees [32]. All these methods aim to improve liveability by using a tool such as object detection to recognise predefined visual elements that the municipality *knows* affects liveability or safety.

The second category of VUA approaches uses methods to directly classify entire panoramic images into labels. This is what is defined as Perceptive VUA, and what we focus on in this paper. As opposed to the former category, these methods do not receive a predisposed notion (i.e., domain knowledge) of what visual elements are relevant for liveability but learn to extract image features which are used to predict a label. This field of research has arisen from the notion that urban perception can be quantified at a large scale and can be used to predict either objective or subjective labels. Objective labels include socio-economic metrics such as mean income [38], population density [2], and housing prices [18]. Subjective labels that have been explored are human annotations regarding perception of panoramic imagery such as beauty [7, 34], perceived safety [27, 28] or liveliness [7] of neighbourhoods.

For our purposes we only focus on Perceptive VUA and not VUA in general, as methods such as finding potholes are applied and built on municipal knowledge. In order to evaluate whether Perceptive VUA has practical use we similarly to prior work use an objective label: we settled on housing prices as it is a quantifiable metric, available at scale, and it relates to liveability making it important for municipal policy-making. The objective nature of housing prices is an important quality as evaluating an explanation is easier to ground when the labels we are trying to predict are non-ambiguous. Inferring how well the model explains why an image is considered beautiful would be harder to evaluate as we as authors might be biased against what is considered beautiful.

2.2 Explainability for Perceptive VUA

Within the field of Perceptive VUA little attention has been given to explainability. [18] utilises computer vision to aid in predicting housing prices alongside a set of multimodal attributes. Their method focuses on the whole image and as such the interpretability of their given desirability score is limited to the scale of their entire image.

A two-step prediction method was used by Gebru et al. [12]. Using an extensively labelled dataset of car brands they were able to predict voting patterns from the cars present in street view imagery. This enabled understanding of what visual element, i.e. brand of car, contributed to the prediction. However, this again only proves that signal exists in this type of data; From a municipal application perspective this is less relevant as for policy-making it is more relevant to find *actionable* visual elements, i.e., elements that can be changed in order to increase the liveability of a neighbourhood. Moreover, from a practical point of view we would like to avoid large scale labelling practices in the form of human annotations as it is a costly process and therefore not accessible for less wealthy cities or countries. Housing prices are already collected by the municipality and thus freely available. A similar note can be made about panoramic imagery, which can be costly to collect and may not be available globally. However, (part of) this cost can be avoided by relying on methodologies for using imagery obtained from Google Street View [26].

An avant la lettre attempt at explainability is introduced by [2] that uses patches from a method by [6] which has been defined as an instance of explaining by examples [16]. This method samples image patches using an HOG representation and trains SVMs to cluster visually similar patches. [2] builds a regressor on top of the resulting SVM bank and thus uses the patches to validate predictive relationships between visual elements and non-visual attributes such as housing prices. A note to the specific details of the approach is to be made. While their approach used visual elements alongside labels such as housing prices, they also used labels such as crime rates. For our purpose it is important to hold back on the invasive nature of AI technology and as such only use labels that we feel are non-harmful. Furthermore, their choice of the housing price label, and as such their analysis, reflected the relation between visual elements and housing prices as is. In our approach the choice of housing prices has been chosen within the scope of liveability and as such will be analyzed within that frame of reference.

A common notion is that traditional methods are more interpretable than Deep Learning methods. Whilst this does not apply to many of the Bag-of-Visual-Words approaches that preceded Deep Learning, we can nonetheless observe that many traditional methods are more straightforward in their execution which aids in interpreting the results, as for instance with the method by [6]. With the rise of black box computer vision the necessity arose for explaining these methods [35]. Initially, methods for interpretation focused primarily on post-hoc explanation, such as GradCAM [33], LIME [29], SHAP [20], and IBA[31]. Whilst these methods moved the field of explainability forward they do not provide direct explanations

of what contributed to the result. This has led to criticism of these methods as the results can be misleading [1, 9, 14]. A proposed alternative involves methods which are self-explaining or inherently explainable [8]. An example of an inherently explainable method is ProtoPNet [3]: a deep network that learns prototypes, that can be grounded in the visual input, and which are directly used in the prediction process. As the explanation is inherent it prevents many of ways in which the explanation can be misleading.

3 EXPERIMENTS FOR PERCEPTIVE VUA

Our goal is to analyse whether the current paradigm of Perceptive VUA has the potential to be used within a municipal setting. From the perspective of policymakers the goal would be to understand the relation between visual elements in street view imagery and socio-economic variables. As most Perceptive VUA methods rely on predicting a label directly we employ explainability methods to generate explanations that can help us understand such a relationship for a socio-economic variable such as housing prices. We choose housing prices as they are objective, available at scale, and capture aspects of liveability which is important for municipal policy-making. Most socio-economic variables are highly correlated and we believe housing prices is a straightforward choice as it has a strong relation to the visual elements visible from street view imagery [2]. In this section we describe the quantitative comparison between different implementations of the Perceptive VUA paradigm including the results, followed by the expert user-study in Section 4.

We use a curated panoramic imagery dataset of Amsterdam [13] in addition to housing prices collected by the municipality covering the entire city. We constrain our scope to the evaluation of three methods that in our view broadly cover the landscape of Perceptive VUA while taking the added dimension of explainability into account: (1) A traditional computer vision method that preceded the black box models focusing on extracting visual elements and as such is considered more interpretable [2, 6], (2) a deep vision method taking the four viewpoint images of the capturing vehicle into account as championed by [38] in combination with a post-hoc explainability technique, (3) and an inherently interpretable model trained through the same deep learning philosophy as [7, 34]. Other methods in the field exists as variations on these methods (i.e. using a different backbone). As such we think these methods adequately cover the existing Perceptive VUA landscape. Furthermore, segmentation-based methods designed to dissect the urban landscape are considered outside of the scope of this research as they are yet to be used for Perceptive VUA.

Patch-Doersch [6]. For our traditional computer vision method we use the discriminative patches approach introduced by Doersch et al. [6]. This method searches for repetitive and discriminative image patches in a binary split image dataset. Patches are represented using HOG descriptors and then clustered into similar looking patches using a bank of SVMs. As there is no clear name given to this method by the authors, we will refer to it as Patch-Doersch. The original code was only available in MATLAB so for the sake of practicality this was implemented in Python. The code will be freely available at github.com/author/Timalph/Patch-Doersch.

Suel-IBA [38]. As the deep vision method we use the model proposed by Suel et al. [38]. This model uses VGG16 [36] as a backbone for extracting features from four directional images of a capturing vehicle. Thereby using the four 4096D feature vectors as input to a parallel fully connected layer aggregating and producing a single value which is used as input for an ordinal loss across all available classes. For our post-hoc explainability method we use IBA [31], which is applied to the deep network to generate a heatmap on top of a processed image using activation mapping. This may allow the user to infer what regions of the image were important for the decision of the model. IBA was found to both qualitatively and quantitatively outperform GradCAM in medical settings [5] and shown to have superior soundness when compared to other XAI benchmarks [19].

ProtoPNet [3]. For the inherently explainable deep network we choose ProtoPNet [3]. This network uses learnt prototypical parts of the image as evidence for making a classification. Their network has built-in case-based reasoning and as such has been described as working similarly to the way human experts make their decisions [3].

For the analysis we focus on Amsterdam, a city with a large diversity in architectural styles between neighbourhoods. To maintain congruity with other approaches within the field of Perceptive VUA we choose to work with the entire city as opposed to a subset. This makes the task significantly harder as models might focus more on geographical elements as a predictor as housing prices in the city centre tend to be higher than in the outskirts. This is shown in Figure 9a in Appendix B. Furthermore, class imbalance was not accounted for, as we are evaluating the task for application in a real municipal setting, as opposed to a lab setting. Furthermore we analyse the practical aspects important for a municipality such as implementation time, running time, and ease of use.

3.1 Datasets and Preprocessing

For our analysis we use a subset of the Amsterdam dataset [13] which consists of 323,124 panoramas taken in 2019 linked to housing prices recorded in 2018. A plot is shown in Figure 9a in Appendix B. All panoramas are oriented to face the front of the capturing vehicle and subsequently bent back to four directional images using equilateral projection. For each panorama in our dataset this results in 4 images of 512x512 pixels of the front, right, back, and left of the capturing vehicle. In the Amsterdam dataset [13] there is metadata of buildings present on the image. Each one of these buildings has an object_id with an associated housing price. Housing prices are averaged for each panorama, while ignoring NaN values and panoramas without buildings. For our purposes the concept of liveability extends to the direct surroundings only. The housing prices are in euros per m^2 and binned according to the bins available on the municipality website[1] as for some housing prices only their respective bin is recorded. The bins are shown in Figure 9b in Appendix B.

3.2 Experimental Setup

Given large differences between the methods used we specify the experimental setup for each method individually:

[1] https://maps.amsterdam.nl/woningwaarde/

Patch-Doersch [6]. The core of this method is centred around learning to distinguish a positive set of images from a negative set, as such the Patch-Doersch method requires a binary dataset. We divided our curated dataset into a split with high housing prices (class 8-11) and a split with low housing prices (class 0-3). The middle classes of 4-7 were ignored to reinforce the binary nature of the dataset. We randomly sample 1000 images for each class, resulting in 4000 images for both the positive and negative set. We run the model in both directions to find patches indicative of high housing prices and vice versa. We use three training iterations for the bank of SVMs.

Suel-IBA [38]. For the post-hoc method we randomly split the curated dataset into a training, validation, and test set 70/20/10. We ensured the classes were divided equally to establish the same class distribution over all sets. We use the deep vision setup as described in [38], and train four parallel fully connected architectures using SGD with a learning rate of 0.001 and a momentum of 0.9. The model was trained for 25 epochs, after which IBA was applied in the form of a Per-Sample Bottleneck to $conv4\_1$ as provided by the authors. The bottleneck was trained separately for all four directions for 5000 iterations.

ProtoPNet [3]. For ProtoPNet we use the same 70/20/10 split as for Suel-IBA. We skip data augmentation used by the authors as we have 1.3M images which is vastly more than the 12,000 images used in the original paper [3], even after they inflate this to 300k images through data augmentation. We do push regression at epoch 11 and 22. The total training time is 23 epochs, which took two weeks. Training was stopped here as training for longer was impractical and outside the scope of our research.

3.3 Experimental Results

In this section we present the results of the comparison between three Perspective VUA methods applied to panoramic data from Amsterdam. For the Patch-Doersch method we clustered repetitive visual elements using HOG descriptors and a bank of SVMs on the binary split dataset, for the Suel-IBA and ProtoPnet methods we used the entire dataset with 12 classes of housing prices. The results inform us what types of patterns we can identify and what they tell us about the relation between urban visual elements and housing prices. In the supplementary material we show additional visual results to complement the results shown here.

Patch-Doersch. Examples of patches identified by the traditional Patch-Doersch approach are shown in Figure 1. Each row depicts a cluster of visual elements. In each row, the first patch is the cluster center. The remaining 7 patches in each row are the closest patches in feature space to the cluster center. The patches for low-housing prices can be seen in Figure 1a. The results of this approach are visually appealing; they show strong visual relationships, but due to the nature of HOG features they primarily capture patterns based on low-level visual features. Some clusters include images that are visually similar but of a different category, such as the red fence in the cluster of flats in the fifth row in Figure 1a. From visual inspection the repeating patterns we can identify are similar types of flats, low-rise houses, certain types of seemingly isolated trees, and empty roads and bicycle paths.

Figure 1b shows the patches for high housing prices. At first glance the areas shown in the patches look more well off. Most patches show an abundance of visual elements which seems to imply liveliness. Another interesting aspect is while the streets are relatively homogeneous in nature, the architecture within buildings seems to be heterogeneous. This might add to the uniqueness of these neighbourhoods. Furthermore, note that in both sets of patches we can identify patterns of parked cars. One could argue that the method has uncovered cheap or expensive cars as discriminative elements, but the distinction appears to be mainly based on the orientation of the cars. For this method the angle at which parking spots are built results in a distinct visual element.

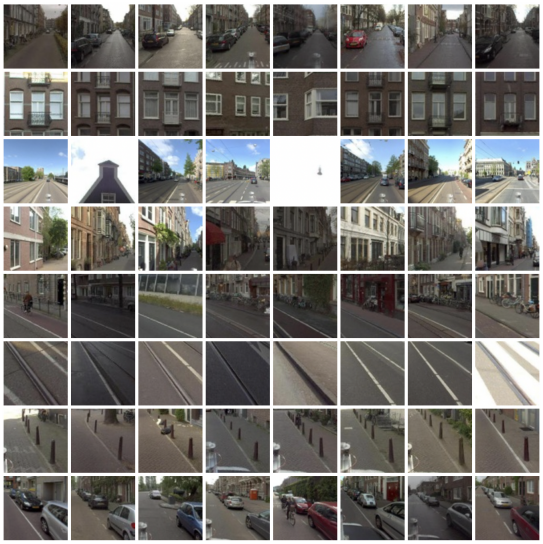
Suel-IBA. Results of the Suel-IBA approach can be seen in Figure 2. The heatmaps have been scaled to the minimum and maximum values of each image individually, represented by the legend on the right of each image. For the results on images of low housing prices in the left column we see patterns focusing on windows, lamp posts, bins, and architecture. For the results on images of high housing prices in the right column we see patterns focusing on architecture, wheels, roofs, and miscellaneous objects.

In general, the heatmaps are spread out in their visualization. Due to the scattered nature of the heatmap it is hard to infer recurring patterns. Furthermore, while we can see that cars, and specifically car tires, tend to elicit a predictive signal it is hard to rely on these elements as it is unclear whether this is spurious correlation or a reliable predictor. The danger of assigning meaning to such a visual element is that a user might be influenced by confirmation bias.

ProtoPNet. The results for ProtoPNet are shown in Figure 3. The left three rows show the prototypes for low housing prices. They contain recurring highrise flats, a lot of green grass next to the road with empty spaces, and a recurring type of low rise architecture with similar small windows. The right three rows show the prototypes for high housing prices: large older style buildings with many windows, combinations of trees along the architecture as seen in the centre of the city, and broad roads with trees next to them for mid-range housing prices areas. However, while semantically similar it is not always clear what the model is highlighting. The boxes are large and often contain multiple visual elements, again making it easy to fall prey to confirmation bias.

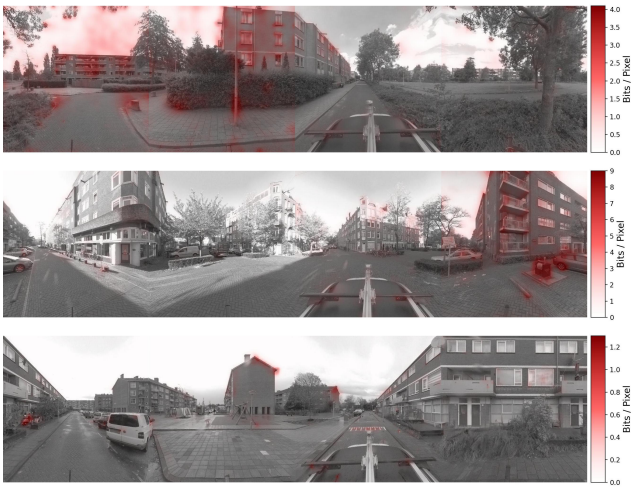
Differences between methods. Both Patch-Doersch and ProtoPNet return patterns that appear more human interpretable. As Patch-Doersch uses low-level features the visual patterns are low-level as well. ProtoPNet identifies visual elements at a higher semantic level, showing similar flats or roofs from different angles. Furthermore, ProtoPNet is not restricted to square image patches, which allows for more freedom in the returned visual elements. Another advantage ProtoPNet has is that the results are visualised in the original image as opposed to cut out and presented without context. This is a design choice that allows the user to immediately ground the explanation in the real world. Comparing these methods to Suel-IBA is hard as the elements highlighted by Suel-IBA are not constrained to be similar. As such the result is not directly interpretable to a human observer, and the patterns cannot be concretely identified. Another important point regarding Patch-Doersch is that the method is restricted to a binary dataset and can therefore only learn patterns for either low or high housing prices. This differs

(a) Patches for lower housing prices (quantile 0-3).



(b) Patches for higher housing prices (quantile 8-11)

Figure 1: Discriminative patches for low and high housing prices based on the Patch-Doersch method.



(a) Suel-IBA output for low housing prices

(b) Suel-IBA output for high housing prices

Figure 2: Bounding box annotations of IBA output. Left column is low housing prices, right is high housing prices. Red values are scaled for min/max values separately for each image.

from both ProtoPNet and Suel-IBA that can learn patterns for all housing price bins.

Quantitative Results The prediction performance of the Suel-IBA and ProtoPNet approaches can be found in Table 1. As the Patch-Doersch approach performs clustering on a binary dataset it has no predictive capability. Important to note is that the Suel-IBA method was trained with an ordinal classification loss, while ProtoPNet was trained with categorical classification. This gives the Suel-IBA method an advantage for the ± 1-2 accuracy. The accuracies show that predicting housing prices purely from visual elements is hard, which may compromise interpretability.

Apart from accuracy we are also interested in practical applicability, i.e., ease of implementation and deployment time. The Suel-IBA approach was the fastest to implement. Preprocessing the VGG16 feature vectors took 45 hours, after which the training time took about 12 hours. All of this was done on a single GPU. ProtoPNet took significantly longer: where training was stopped after training for two weeks on 4 GPUs. The traditional Patch-Doersch approach took about 60 hours. Note that the majority of this model runs on a CPU, utilising a single GPU only for computing nearest neighbours. It also has the large number of hyperparameters that due to the long

(a) ProtoPNet prototypes for lower housing prices. **(b) ProtoPNet prototypes for higher housing prices.**

Figure 3: Results from ProtoPNet. Closest neighbours to each prototype in the left column. The top three rows are prototypes found for low-housing prices. The bottom three rows are prototypes found for high-housing prices.

training time and lack of quantitative evaluation were impractical to explore.

Models	±0 Acc	±1 Acc	±2 Acc
Patch-Doersch	-	-	-
SuelIBA	.32	.70	.86
ProtoPNet	.16	.33	.47

Table 1: Prediction accuracy for the class labels. Patches method has no predictive capability. ± 1-2 indicates accuracy within 1-2 labels of the correct label.

4 SURVEY STUDY

To evaluate the usability of the Perceptive VUA paradigm for municipal purposes we performed an Expert User-Study. For this we approached researchers from the UvA (University of Amsterdam) in Computer Vision and Responsible AI, and employees from the municipality that have a relevant AI/Data Science background and asked them to fill out a survey that compares the three Perceptive VUA methods.

4.1 Participants

In order to get a representative group of respondents we built separate populations for each participant group: employees at the municipalities of Amsterdam and Rotterdam, Responsible AI researchers and Computer Vision researchers. As there are only a select amount of employees at the municipality that have a relevant background we approached participants directly. They are all employees that work either in policy-making regarding new technology, or work in Data Science or Artificial Intelligence at the municipality. As such we approached 7 employees that work in the tech innovation department, who's job it is to assess whether new technological innovations can be implemented around the city. We also approach 13 employees from AI, Computer Vision, and Data Science teams that work on AI, Computer Vision, and Data Science solutions specifically for the city. For the Responsible AI researchers we built a population consisting of all researchers working directly for, or in labs associated with, the UvA who's primary research focus is on explainability, trustworthiness, fairness, privacy, or transparency in AI. The population consisted of 49 researchers of which we sampled 15 participants. Finally, our population of Computer Vision researchers consisted of all researchers working directly for, or in labs associated with, the UvA who's primary research focus revolves around using Computer Vision techniques. The population consisted of 70 researchers out of which we sampled 18 participants. Informative consent was received orally before taking the survey in addition to written consent which was received after. All surveys were anonymous and compensation in the form of chocolate was offered but not always accepted. This resulted in 51 valid responses: 19 municipal employees, 18 CV researchers, and 14 Responsible AI researchers. 2 incomplete responses were discarded.

4.2 Materials and Procedure

At the start of the survey the participants are told that the housing price bins are referred to with textual equivalents: Lowest, Low, Average, Above Average, High or Very High. This to de-emphasise the actual prices and to reduce confusion while evaluating the explanations. The survey consists of 5 sections in order. A visualisation can be seen in Figure 4. The survey can be viewed in its entirety in Appendix C.

(1) Section 1 consists of a primer of 5 questions where participants are presented with two panoramic images of Amsterdam neighbourhoods and are required to pick the one with the highest average housing prices. This to evaluate whether they are capable of judging the relationship between visual elements and housing prices. The images are selected by the first author with the criteria of covering a variety neighbourhoods and housing types. Where possible we tried to control for lighting and weather conditions. Both the order of the 5 questions as the order of the two images presented are randomized.
(2) Section 2 is an evaluation of each method separately in which we show each of the three methods in sequence, and present

the participant with six explanations per method: 3 for higher housing prices, and 3 for lower housing prices. The methods are shown in a random order. The explanations within each method are also randomised. Each explanation consists of two sub-questions. The first aims to evaluate whether the visual elements used by the method were the same as a human would use for this task: *I personally would have used the same visual elements for this decision.* This, because it has been shown that human reasoning is an important quality for real world XAI applications [17]. The second question aims to evaluate whether based on the returned explanation the participant thinks it makes sense to rely on the AI's decision: *Based on these visual elements, I think it makes sense to rely on the AI's decision.* This, because in the public sector reliability is important within the context of AI safety in order to minimize potential downstream harm to citizens [40]. These statements are then rated on a five-point Likert scale. The order of these two sub-questions was determined randomly at the start of a survey and then kept consistent throughout.

The 6 explanations are then followed by a summative question in which all the explanations are shown at the same time alongside the summative question on how much the participant trusts the model's understanding of the relationship between visual elements and housing prices.

(3) Section 3 consists of comparative questions; four multiple choice questions comparing the methods on a similar scene where each choice is an explanation. We do this by generating explanations for a single housing price bin for each of the methods. We do this four times, and ask: *Which of these methods does the best job at returning visual elements that you would use for this decision?*
(4) Section 4 consists of utility questions whether the participant would use the methods for their work. This to measure the practical use of these methods. The order of the methods is randomized.
(5) In section 5 we asked demographic questions, including ranking their knowledge of Artificial Intelligence and their knowledge of Computer Vision separately. Furthermore we ask whether they work at a municipality and if they do, to what extent they are involved in policy-making and to what extent their work is technical or non-technical. If they do not work at the municipality, we ask them to rate how much their research focuses on Responsible AI. All questions are evaluated on a scale of 1-10, apart from the policy-making question which is multiple choice.

4.3 Analysis

For the primer, in which the participants were asked to select the picture containing a neighbourhood with higher average housing prices, the results were as follows: 31 participants made 0 mistakes, 15 participants made 1 mistake, and 5 participants made 2 mistakes. As all participants made less than 50% mistakes we included all responses in our results. For section 2 we aggregated the results per method by summing over all explanations. These are shown in Figure 5. Our directional hypotheses are that Patch-Doersch scores most favourably, followed by ProtoPNet, and lastly Suel-IBA. As our data does not follow a normal distribution we perform pairwise right tailed Wilcoxon signed-rank tests. Bonferroni correction is applied for all pairwise tests with $m = 3$ and as such our null hypotheses will be rejected if $p < 0.01667$. For the trustscores we test the ranking using pairwise one-sided independent t-tests. Our directional hypotheses are that Patch-Doersch has the highest trustscore, followed by ProtoPNet, and lastly Suel-IBA. Our null hypotheses will be rejected if $p < 0.01667$. For the section 3 comparative questions, we aggregate the responses by computing the mean value and calculate confidence intervals using the margin of error. For the section 4 utility questions regarding the use of the methods in participants' daily work we test the directional hypotheses that Patch-Doersch scores most favourably, followed by ProtoPNet, and lastly Suel-IBA. We do pairwise right tailed Wilcoxon signed-rank tests with Bonferroni correction for $m = 3$ and as such reject our null hypotheses if $p < 0.01667$.

4.4 Results

In Figure 5 we see the summed responses for the individual explanations for each method. The first thing to note is that we see that most responses do not extend to the strongly agree or strongly disagree areas. Of the three methods, along the two dimensions, Patch-Doersch seems to have elicited the strongest response, with the largest portion of the answers leaning to the right side of neutral. The next most favourable seems to be ProtoPNet, followed by Suel-IBA. We tested this ranking using pairwise right tailed Wilcoxon signed rank-tests and observed all three p-values to be < 0.001. As such the results favoured the alternative hypotheses and confirm this ranking. More in-depth statistics regarding tests are shown in Table 2 in Appendix A.

The participants often have contradicting opinions when shown the individual explanations, but for some cases there is positive consensus; More than two-thirds of participants vote to the right side of neutral on both subquestions. The three explanations that satisfied this are all by Patch-Doersch. These explanations also received the highest percentage of *Strongly Agree* responses. They are shown in Figure 6, and show expensive houses in the city centre, windows belonging to expensive housing, and cheaper flats.

The obtained trustscores are shown on the left in Figure 7 and show a similar picture. The mean trustscores are 5.67 (Patch-Doersch), 4.16 (Suel-IBA), and 4.84 (ProtoPNet). The ranking of Patch-Doersch > ProtoPNet > Suel-IBA can not be confirmed however, as only the p-value for Patch-Doersch > Suel-IBA is below our Bonferroni corrected α. The p-values for the comparisons between the trustscores of different methods are shown in Table 2 in Appendix A alongside the CI's and Cliff's Delta.

The result and confidence intervals for the comparison of methods on the question *Which of these methods does the best job at returning visual elements that you would use for this decision* can be viewed on the right in Figure 7. Patch-Doersch is most often chosen as the method that does the best job.

In Figure 8 we see that no methods would be directly usable. We again tested the pairwise ranking using right tailed Wilcoxon signed rank-tests and observed p-values < 0.001, thus we accept

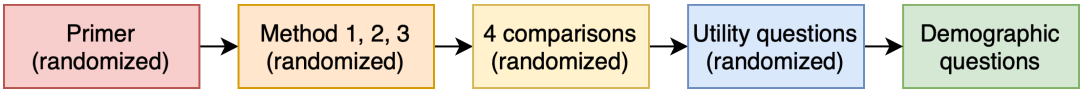


Figure 4: Survey Study Procedure. Informed consent was received orally prior to the survey and written post-survey.

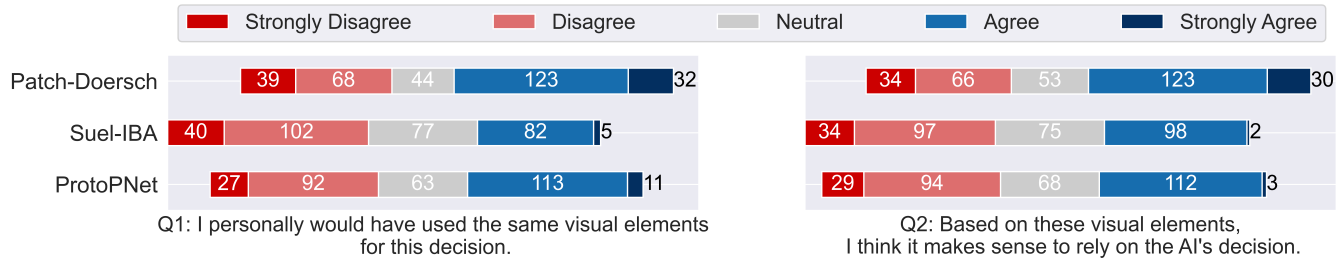


Figure 5: Total summed responses for survey of individual explanations for the two statements in section 2. Bars add up to n=(306).

(a) Patch-Doersch explanation for high housing prices. Q1: 55% Agree, 27% S. Agree. Q2: 63% Agree, 19% S. Agree.

(b) Patch-Doersch explanation for low housing prices. Q1: 53% Agree, 17% S. Agree. Q2: 53% Agree, 14% S. Agree.

(c) Patch-Doersch explanation for high housing prices. Q1: 57% Agree, 18% S. Agree. Q2: 57% Agree, 12% S. Agree.

Figure 6: Three individual explanations with the highest amount of participants responding to the right side of neutral for both section 2 subquestions: *"I personally would have used the same visual elements for this decision"* and *"Based on these visual elements, I think it makes sense to rely on the AI's decision"*. Q1 and Q2 refer to the order as presented in Figure 5.

the alternative hypothesis for the ranking of Patch-Doersch > ProtoPNet > Suel-IBA. This, together with the favourability for the individual explanations, the trustscores given by the experts for each method and the comparisons in Figure 7 show us that of the three methods, Patch-Doersch is most often favoured. The self-reported demographics for section 5 are shown in Table 10 in in Appendix B. Further statistics for each method are shown in Table 3 in Appendix B.

Textual responses showed certain nuances regarding the less favourable view of Suel-IBA. Participants responded that the method tended to highlight irrelevant features which lowered the trust in the model.

"*It focuses on very few and/or small areas. It also has a* ***tendency to focus on the car the camera was on*** *which was in many of the examples, it also* ***highlights random parts*** *of the street, sky, or facade* ***that make it untrustworthy***" "*Things like* ***parts of the camera car are taken into account*** *when they clearly shouldn't be. This* ***lowers my trust in the model***." "***This heatmap 'looks' far to much at*** *car windows, tires, bicycles, poles and other* ***non-relevant elements***. *The only heatmap I thought was slightly passable was the one that coloured bell gables red. (I would personally look at that too)." - Translated from Dutch. "This method seems* ***super focused on cars***."

Consistency is consistently mentioned as an important factor for viewing a method positively. ProtoPNet is often accused of returning patterns of dissimilar visual elements. "*I noticed that* ***a single weird example*** *in the list [returned by Patch-Doersch]* ***would throw me off***" "*The [Patch-Doersch] images clearly show the same visual aspects (****it's consistent****, unlike [ProtoPNet] in my opinion), which* ***gives a sense of certainty*** *of the model prediction.*" "*[ProtoPNet] seems reasonable but again* ***not very consistent***. "*I am sometimes confused as* ***the box*** *highlighted [by ProtoPNet], to my eye, [it]* ***does not show*** *any* ***similarity to the predicted picture***." "***The selection*** *of the visual element [ProtoPNet] focus on* ***it's very broad***." Multiple

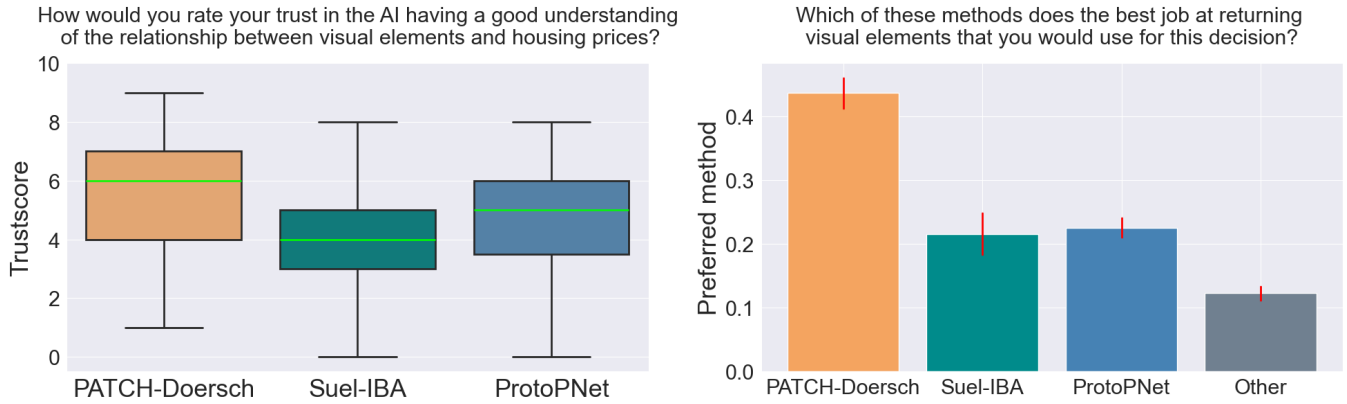


Figure 7: Left: After six examples for a method respondents were presented with an overview of all the examples of that method and asked: How would you rate your trust in the AI having a good understanding of the relationship between visual elements and housing prices? With 10 meaning you trust the AI completely, and 0 meaning you do not trust the AI at all. The green line shows the median. Right: Percentual responses for the comparison of explanations run within a certain neighbourhood/housing price group for the question: *Which of these methods does the best job at returning visual elements that you would use for this decision?*

participants mentioned concerns regarding the uncertainty that comes with linking their symbolic interpretation to the interpretation given by the model. This is a pitfall for confirmation bias. "*Mostly, in these [Suel-IBA] examples, I am **not sure what part is highlighted** and/or **don't see the relevancy** of it. Even if a sensible part of the image is highlighted, I wouldn't know why if the **meaning I attribute** to it is the **same as what the model detects**, and I may **fall prey to confirmation bias**.*" "*I like this method. However, it is **not straightforward to link** the continuous type of interpretation (**pixel intensity**) **to** our more **symbolic interpretation**. This is to say that this method is somehow **prone to confirmation bias**. None*" "*again, **some boxes are drawn**. I've **no way for verifying what they are / mean**.*" "*i don't know/understand what those red areas [returned by Suel-IBA] are; there is **no way to verify what that means** or to verify the veracity of the highlighted regions.*

Finally, missing context and the naive approach for the tasks led to some participants mentioning the methods need to be improved. "*I think that for detecting a price of the house in Amsterdam **neither of those methods** alone **are sufficient**.*" "*I generally like [Patch-Doersch], but it **does not consider the context surrounding the houses**, which I find a **crucial aspect for this task** (at least important for me when performing it).*" "*[ProtoPNet] **missing social economic en cultural context**. AI is **not precise enough** at this moment.*"

5 DISCUSSION

Based on the results we obtained we can make a number of observations. The Patch-Doersch approach is clearly the most favoured for giving insight into the relation between visual elements and housing prices, both from our own perspective as from the rankings of the methods uncovered through the survey. This is reflected in both the returned visual elements as the trust scores, and as such the Patch-Doersch approach seems to be the closest to municipal use. ProtoPNet scores less well, as it often confuses visual elements that are not similar or highlights too broad a selection making it unclear what visual elements the explanation contains. The Suel-IBA approach returned heatmaps highlighting both relevant and irrelevant elements thus making it hard to trust due to potential confirmation bias. Our survey confirms this both through textual responses as through the trustscore, which on average was the lowest. More than half of the experts also indicated they would not use this. The lack of identifiable patterns is partly to be expected, as there is no mechanism in IBA to learn these patterns. As a method for supporting municipalities we can thus say both ProtoPNet and Suel-IBA are not adequate.

While the Patch-Doersch approach does produce patterns, it has a number of limitations that need to be taken into account. First off it demands a binary dataset. This is because it was originally developed to find repetitive visual elements for a single city (i.e., "*What makes Paris look like Paris?*"), with multiple other cities in the negative set. When trained, Patch-Doersch often finds visual elements predictive of a neighbourhood, instead of predicting the socio-economic dimension. The visual elements in most of the returned clusters were within the same neighbourhood, as housing prices are very much geographically distributed in Amsterdam. While these might still be indicators of housing prices, we assume that stronger indicators might exist that are now being obscured by geographical indicators in the form of visual elements. Secondly, it has no predictive capability. As such all clusters are visualizations of training data. A predictor has been built on top of it by [2], but due to training times this was infeasible for our approach. Patch-Doersch is also relatively slow making it hard to run many experiments. A single training run on less than 10k images takes 60 hours for three SVM iterations. This makes it infeasible to run a large dataset. The dataset used by Patch-Doersch is a fraction of the 300k, and 323k images ProtoPNet and Suel-IBA are respectively trained on, and as such cannot be viewed as reliable representation of the city. Lastly, the low-level HOG features used in this method are not rotationally invariant and do not describe any semantics of the patches. As

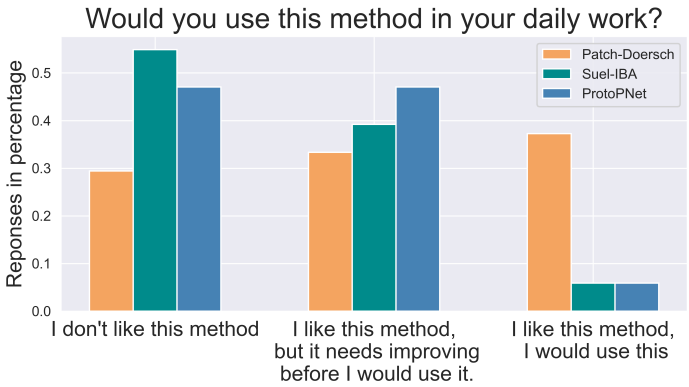


Figure 8: Results for the question whether a participant would use this in their daily work. Percentages for the three *I don't like this method* options are summed.

such, the clusters are heavily influenced by camera orientation and low-level patterns.

ProtoPNet runs relatively fast, albeit on multiple GPUs, but does not scale well to the entirety of Amsterdam. This is to be expected, as the dataset of roughly 1.2M images is much larger than the 300k images ProtoPNet was trained on. The visual elements returned by ProtoPNet are more sophisticated than those by the Patch-Doersch approach, but they are still naively informed visual elements. While the method could use improving, the approach is more suitable than a standard deep vision model. Ideally, we would like to deconstruct a panorama and understand the visual elements that relate to the socio-economic variable, something that ProtoPNet at the very least attempts to do.

Finally, all predictions shown to respondents were correct predictions made by the model. As mentioned before, for Patch-Doersch these were clusters it had been trained on. As such these were the best possible examples for each method. Of the possible socio-economic variables that have previously been used in Perspective VUA settings, housing prices is low-hanging fruit as opposed to metrics such as, education rates [38], scenicness [34], or theft [2] or crime rates [11]. With that in mind, one could argue that the moderate trustscores recorded in the survey, with all methods averaging below 6, does not make for viable methods in practice. Further discussion on limitations and recommendations for municipalities can be found in Appendix A.

6 CONCLUSION

We set out to evaluate whether the current paradigm of Perceptive VUA is suitable yet for policymakers at a municipality. For this we compared three different Perceptive VUA methods in combination with explainability techniques for analysing the relationship between visual elements and housing prices in the city of Amsterdam. In the results we saw that the Patch-Doersch approach was able to provide recurring visual patterns, while the ProtoPNet and Suel-IBA were inconsistent and therefore less trustworthy. We evaluated the methods with experts from the Municipality of Amsterdam and Rotterdam, as well as Responsible AI and Computer Vision researchers and found that none of the models met the trustworthiness criteria. We make recommendations for a model to explain urban visual relationships including both practical dimensions and explainable capabilities. Overall, none of the methods explored meet the criteria yet to successfully support policy-making in a municipal context. This is confirmed by the experts. Patch-Doersch comes closest, and in our opinion shows the direction the Perceptive VUA paradigm needs to move in: a supporting tool that returns consistent and trustworthy explanations. As such we believe our conclusion to be a hopeful one: Perceptive VUA has the potential to be a powerful tool for policymakers, but in order for that to happen the focus needs to shift from exploring theoretical capabilities to meeting the needs of the people it is supposedly designed for.

7 ETHICS

Urban Analytics in its current state is an inherently sensitive subject. Predicting socio-economic variables through a representation of a neighbourhood in the form of panoramic imagery might be possible when evaluated purely on an accuracy metric, but policy built on this could suffer from turning a blind eye to individual harm exerted by such an application. As these applications revolve around categorizing areas people live, there is potential for downstream harm to citizens.

At first glance harmless objectives, such as uncovering the relationship between visual elements and housing prices, can have undesired side effects if methods can not be trusted to identify concrete and trustworthy visual elements, as opposed to spurious correlations. As our conclusion is that at this time methods for perceptive VUA are not sophisticated enough to produce results that can reliably be used for policy-making, we would urge municipalities to hold-off on implementing such methods.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in Neural Information Processing Systems* 2018-Decem, NeurIPS (2018), 9505–9515. https://doi.org/10.48550/arXiv.1810.03292

[2] Sean M Arietta, Alexei A Efros, Ravi Ramamoorthi, and Maneesh Agrawala. 2014. City forensics: Using visual elements to predict non-visual city attributes. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2624–2633. https://doi.org/10.1109/TVCG.2014.2346446

[3] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. This looks like that: Deep learning for interpretable image

recognition. *Advances in Neural Information Processing Systems* 32, NeurIPS (2019), 1–12. arXiv:1806.10574

[4] Weslei Gomes de Sousa, Elis Regina Pereira de Melo, Paulo Henrique De Souza Bermejo, Rafael Araújo Sousa Farias, and Adalmir Oliveira Gomes. 2019. How and Where Is Artificial Intelligence in the Public Sector Going? A Literature Review and Research Agenda. *Government Information Quarterly* 36, 4 (Oct. 2019), 101392. https://doi.org/10.1016/j.giq.2019.07.004

[5] Ugur Demir, Ismail Irmakci, Elif Keles, Ahmet Topcu, Ziyue Xu, Concetto Spampinato, Sachin Jambawalikar, Evrim Turkbey, Baris Turkbey, and Ulas Bagci. 2021. Information Bottleneck Attribution for Visual Explanations of Diagnosis and Prognosis. In *Machine Learning in Medical Imaging (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 396–405. https://doi.org/10.1007/978-3-030-87589-3_41

[6] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A. Efros. 2015. What makes Paris look like Paris? *Commun. ACM* 58, 12 (2015), 103–110. https://doi.org/10.1145/2830541

[7] A Dubey, N Nikhil, D Parikh, R Raskar, and César A. Hidalgo. 2016. Deep Learning the City: Quantifying Urban Perception at a Global Scale. *Eccv* 3 (2016), 398–413. https://doi.org/10.1007/978-3-319-46448-0

[8] Daniel C. Elton. 2020. Self-explaining ai as an alternative to interpretable ai. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12177 LNAI (2020), 95–106. https://doi.org/10.1007/978-3-030-52152-3_10 arXiv:2002.05149

[9] Hidde Fokkema, Rianne de Heide, and Tim van Erven. 2023. Attribution-based Explanations that Provide Recourse Cannot be Robust. *Journal of Machine Learning Research* 24, 360 (2023), 1–37. http://jmlr.org/papers/v24/23-0042.html

[10] Matt Franchi, J.D. Zamfirescu-Pereira, Wendy Ju, and Emma Pierson. 2023. Detecting disparities in police deployments using dashcam data. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. ACM. https://doi.org/10.1145/3593013.3594020

[11] Kaiqun Fu, Zhiqian Chen, and Chang Tien Lu. 2018. StreetNet: Preference learning with convolutional neural network on urban crime perception. *GIS: Proceedings of the ACM International Symposium on Advances in Geographic Information Systems* August 2019 (2018), 269–278. https://doi.org/10.1145/3274895.3274975

[12] Timnit Gebru, Jonathan Krause, Yilun Wang, Duyun Chen, Jia Deng, Erez Lieberman Aiden, and Li Fei-Fei. 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 (2017), 13108–13113. https://doi.org/10.1073/pnas.1700035114

[13] Inske Groenen, Stevan Rudinac, and Marcel Worring. 2022. PanorAMS: Automatic Annotation for Detecting Objects in Urban Context. (2022). arXiv:2208.14295

[14] Sara Hooker, Dumitru Erhan, Pieter-jan Kindermans, Been Kim, and Google Brain. 2019. A Benchmark for Interpretability Methods in Deep Neural Networks. NeurIPS (2019). arXiv:1806.10758v3

[15] Mohamed R. Ibrahim, James Haworth, and Tao Cheng. 2020. Understanding cities with machine eyes: A review of deep computer vision in urban analytics. *Cities* 96 (2020), 102481. https://doi.org/10.1016/j.cities.2019.102481

[16] Atsushi Kanehira and Tatsuya Harada. 2019. Learning to explain with complemental examples. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2019), 8595–8603. https://doi.org/10.1109/CVPR.2019.00880 arXiv:1812.01280

[17] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM. https://doi.org/10.1145/3544548.3581001

[18] Stephen Law, Brooks Paige, and Chris Russell. 2019. Take a look around: Using street view and satellite images to estimate house prices. *ACM Transactions on Intelligent Systems and Technology* 10, 5 (2019), 1–19. https://doi.org/10.1145/3342240 arXiv:1807.07155

[19] Yawei Li, Yang Zhang, Kenji Kawaguchi, Ashkan Khakzar, Bernd Bischl, and Mina Rezaei. 2023. A Dual-Perspective Approach to Evaluating Feature Attribution Methods. *arXiv preprint* (2023). arXiv:2308.08949

[20] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* Section 2 (2017), 4766–4775. arXiv:1705.07874

[21] Kevin Lynch. 1964. *The image of the city*. The MIT press.

[22] Nachuan Ma, Jiahe Fan, Wenshuo Wang, Jin Wu, Yu Jiang, Lihua Xie, and Rui Fan. 2022. Computer Vision for Road Imaging and Pothole Detection: A State-of-the-Art Review of Systems and Algorithms. (2022), 1–16. https://doi.org/10.1093/tse/tdac026 arXiv:2204.13590

[23] Shannon Mattern. 2017. A City Is Not a Computer. *Places Journal* (2017).

[24] Patrick Mikalef, Siw Olsen Fjørtoft, and Hans Yngvar Torvatn. 2019. Artificial Intelligence in the Public Sector: A Study of Challenges and Opportunities for Norwegian Municipalities. In *Digital Transformation for a Sustainable Society in the 21st Century (Lecture Notes in Computer Science)*. Springer International Publishing, Cham, 267–277. https://doi.org/10.1007/978-3-030-29374-1_22

[25] Slava Jankin Mikhaylov, Marc Esteve, and Averill Campion. 2018. Artificial Intelligence for the Public Sector: Opportunities and Challenges of Cross-Sector Collaboration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376, 2128 (2018). https://doi.org/10.1098/rsta.2017.0357

[26] Emily Muller, Emily Gemmell, Ishmam Choudhury, Ricky Nathvani, Antje Barbara Metzler, James Bennett, Emily Denton, Seth Flaxman, and Majid Ezzati. 2022. *City-Wide Perceptions of Neighbourhood Quality using Street View Images*. Vol. 1. Association for Computing Machinery. arXiv:2211.12139

[27] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and Cesar Hidalgo. 2014. Streetscore-predicting the perceived safety of one million streetscapes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* January (2014), 793–799. https://doi.org/10.1109/CVPRW.2014.121

[28] Vicente Ordonez and Tamara L. Berg. 2014. Learning high-level judgments of urban perception. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 8694 LNCS, PART 6 (2014), 494–510. https://doi.org/10.1007/978-3-319-10599-4_32

[29] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1135–1144. https://doi.org/10.1145/2939672.2939778 arXiv:1602.04938

[30] Tina Ringenson, Mattias Höjer, Anna Kramers, and Anna Viggedal. 2018. Digitalization and Environmental Aims in Municipalities. *Sustainability* 10, 4 (2018), 1278. https://doi.org/10.3390/su10041278

[31] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the Flow: Information Bottlenecks for Attribution. *8th International Conference on Learning Representations* (2020). arXiv:2001.00396

[32] Ian Seiferling, Nikhil Naik, Carlo Ratti, and Raphäel Proulx. 2017. Green streets - Quantifying and mapping urban trees with street-level imagery and computer vision. *Landscape and Urban Planning* 165, 4 (2017), 93–101. https://doi.org/10.1016/j.landurbplan.2017.05.010

[33] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (2020), 336–359. https://doi.org/10.1007/s11263-019-01228-7 arXiv:1610.02391

[34] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. 2017. Using deep learning to quantify the beauty of outdoor places. *Royal Society Open Science* 4, 7 (2017). https://doi.org/10.1098/rsos.170170

[35] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep inside convolutional networks: Visualising image classification models and saliency maps. *2nd International Conference on Learning Representations, Workshop Track Proceedings* (2014), 1–8. arXiv:1312.6034

[36] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations* (2015), 1–14. arXiv:arXiv:1409.1556v6

[37] Esra Suel, Samir Bhatt, Michael Brauer, Seth Flaxman, and Majid Ezzati. 2021. Multimodal deep learning from satellite and street-level imagery for measuring income, overcrowding, and environmental deprivation in urban areas. *Remote Sensing of Environment* 257 (2021), 112339. https://doi.org/10.1016/j.rse.2021.112339

[38] Esra Suel, John W. Polak, James E. Bennett, and Majid Ezzati. 2019. Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports* 9, 1 (2019), 1–10. https://doi.org/10.1038/s41598-019-42036-w

[39] Maarten Sukel, Stevan Rudinac, and Marcel Worring. 2020. Urban object detection kit: A system for collection and analysis of street-level imagery. *Proceedings of the 2020 International Conference on Multimedia Retrieval* (2020), 509–516. https://doi.org/10.1145/3372278.3390708

[40] Bernd W Wirtz, Jan C Weyerer, and Carolin Geyer. 2019. Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration* 42, 7 (2019), 596–615.

[41] Tan Yigitcanlar, Duzgun Agdas, and Kenan Degirmenci. 2023. Artificial Intelligence in Local Governments: Perceptions of City Managers on Prospects, Constraints and Choices. *AI & SOCIETY* 38, 3 (2023), 1135–1150. https://doi.org/10.1007/s00146-022-01450-x

[42] Yonglin Zhang, Shanlin Li, Rencai Dong, Hongbing Deng, Xiao Fu, Chenxing Wang, Tianshu Yu, Tianxia Jia, and Jingzhu Zhao. 2021. Quantifying physical and psychological perceptions of urban scenes using deep learning. *Land Use Policy* 111 (2021), 105762. https://doi.org/10.1016/j.landusepol.2021.105762

[43] Junhan Zhao, Xiang Liu, Yanqun Kuang, Yingjie Victor Chen, and Baijian Yang. 2018. Deep CNN-Based Methods to Evaluate Neighborhood-Scale Urban Valuation Through Street Scenes Perception. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. 20–27. https://doi.org/10.1109/DSC.2018.00012

A APPENDIX

A.1 Recommendations for VUA with application in policy-making

Based on our experiments and evaluation with experts we believe the current paradigm for Perceptive VUA is inherently flawed. This is due to three problems that are currently not addressed by the existing methods:

The current paradigm for Perceptive VUA revolves around brute-forcing a distribution over the urban visual landscape. However, cities are complex entities[23] that do not exist in the same theoretical setting as most computer vision tasks do. There are complex causal relations between socio-economic factors and what we see on the street. As such, visual elements returned by the explanations are often spurious correlations as observed in our qualitative evaluation as well as through the textual survey responses. This is an inherent problem with the current paradigm of Perceptive VUA: The visual nature of the city allows for 'shortcuts' for predictive models in the form of spatially distributed visual elements that do not necessarily have a causal relation with the predicted variable.

Secondly, current methods are not built with explainability, trustworthiness, or domain knowledge in mind. With the exception of Patch-Doersch, the existing paradigm focuses solely on accuracy. While the accuracy of these predictive models is high, the act of condensing the entire urban visual landscape into a single predictive feature in turn makes the method inexplicable and untrustworthy. As mentioned by multiple experts in the survey, trust is extremely important. Municipalities need to be able to rely on the predictions or explanations generated by the AI as they potentially affect real people.

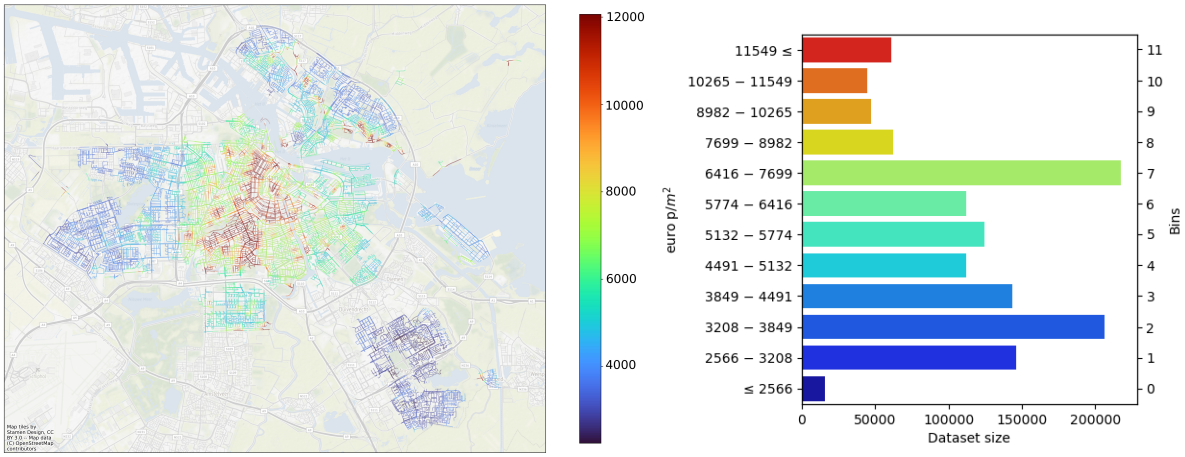
Finally through the textual survey responses we observed municipal employees, and other participants, kept evaluating the methods by relying on their knowledge of the city. Municipal employees *know* the city and as such look for practical tools that can support their knowledge as opposed to a one-size-fits-all model. It is our observation that the current paradigm of Perceptive VUA attempts to serve as an oracle as opposed to a tool; This paradigm is too focused on the nature of learning patterns in big data instead of grounding the models in domain knowledge. Cities are complex, and an approach that takes a set of images and labels as a representation grossly underestimates that complexity and attempting to solve this by teaching an AI socio-economic and cultural context is a solution that will bring even more problems than already exist. As such, we believe the paradigm of Perceptive VUA needs to shift to becoming more of a tool. While Patch-Doersch is not useable due to it's practical and technical drawbacks, we think the relatively positive survey outcome highlights the direction the paradigm of Perceptive VUA needs to move in: A method focused on *supporting* municipalities that provides consistent and trustworthy results where our symbolic interpretation is the same as meaning provided by the AI's prediction. We present the lessons we learned as a recommendation for further research in this field:

- The method, and explainability technique, should be grounded in domain knowledge as we are interested in explanations regarding the real world, not just visually similar images.
- The approach should focus on trustworthy and consistent results, ideally by dissecting the image into single interpretable visual elements.
- The focus should lie on developing tools for experts. One should always ask themselves: How would an expert use this tool? To what extent does an expert need to trust the model to be correct?

A.2 Limitations

We acknowledge our study has certain limitations. Our use of a single label and single city limits the scope of this research such that these results might not be generalisable when applied to a new city. However, we would argue that these methods need to be evaluated by municipalities before they are implemented in a local context. Secondly, while the use of a single label can be viewed as a limiting factor, our choice for housing prices is motivated by the high correlation between most socio-economic labels. We considered housing prices to be reliable as a variable as it does more directly relate to the visual elements visible from street view imagery [2] than a label such as unemployment [38]. We argue that if it is not possible for a Perceptive VUA method to correctly identify relevant visual elements for an 'easy' label such as housing prices, it will be less reliable or useful for a hard and potentially problematic label such as unemployment. Finally, our claim is not to be read as an assessment of VUA or Perceptive VUA in general, but an assessment about Perceptive VUA in municipal contexts only. We argue that this recipe, or paradigm, of naively training models to directly map urban imagery to a socio-economic variable is, as of now, insufficient to be implemented within a municipal context.

B APPENDIX



(a) Map of Amsterdam with a heatmap overlay of the housing prices in euro p/m^2 in 2018.

(b) Housing price bins as provided by the municipality website.

Figure 9: Map of housing prices alongside housing price bins. Note how the distribution of high housing prices is clustered around the centre, with housing prices declining gradually as we move further towards the edge of the city.

	#	Kn. of AI	Kn. of CV	Research focused on resp. AI	Kn. of policy-making within municipality	Technical nature of day-to-day work
Municipal employees	19	6.7 ± 2.6	6.4 ± 1.8	X	6.1 ± 2.4	6.6 ± 2.3
CV Researchers	18	8.2 ± 0.9	8.1 ± 0.7	5.4 ± 3.1	X	X
Resp. AI researchers	14	7.9 ± 1.8	5.4 ± 2.1	8.3 ± 1.6	X	X

Figure 10: Self-reported survey demographics, columns show mean ± sd.

Right tailed Wilcoxon signed rank-tests *I personally would have used the same visual elements for this decision*			
	Patch-Doersch > Suel-IBA	Patch-Doersch > ProtoPnet	ProtoPNet > Suel-IBA
p-value	**< 0.001***	**< 0.001***	**< 0.001***
Test statistic	8460	2464	3160
Cliff's Delta	0.209	0.091	0.132
Sample size	306	306	306
Based on these visual elements, I think it makes sense to rely on the AI's decision			
p-value	**< 0.001***	**< 0.001***	**< 0.001***
Test statistic	6328	4136	435
Cliff's Delta	0.185	0.139	0.050
Sample size	306	306	306
Would you use this method in your daily work?			
p-value	**< 0.001***	**< 0.001***	**< 0.001***
Test statistic	780.0	55.00	528.0
Cliff's Delta	0.384	0.326	0.086
Sample size	51	51	51

One-sided independent t-tests *Differences in trustscores*			
	Patch-Doersch > Suel-IBA	Patch-Doersch > ProtoPnet	ProtoPNet > Suel-IBA
p-value	**0.0003***	0.0308	0.0384
Test statistic	3.547	1.890	1.788
Cliff's Delta	3.991	0.262	0.220
CI	1.5 ± 0.84	0.82 ± 0.86	0.686 ± 0.75
Sample size	51	51	51

Table 2: P-values for right tailed Wilcoxon signed rank-tests for responses to section 2 (individual explanations) and section 4 (utility questions) and p-values for one-sided independent t-tests for the differences in trustscores. After Bonferroni correction $\alpha = 0.01667$. Statistically significant results are denoted in bold font with an asterisk.

	Patch-Doersch	Suel-IBA	ProtoPNet
I personally would have used the same visual elements for this decision	4 ± 1 (n=306)	3 ± 1 (n=306)	3 ± 1 (n=306)
Based on these visual elements, I think it makes sense to rely on the AI's decision	3.5 ± 1 (n=306)	3 ± 1 (n=306)	3 ± 1 (n=306)
Trustscores	5.67 ± 2.4 (n=51)	4.16 ± 1.9 (n=51)	4.84 ± 2.0 (n=51)
Would you use this method in your daily work?	4 ± 1 (n=51)	3.0 ± 1.5 (n=51)	4.0 ± 1.0 (n=51)

Table 3: Medians and semi-iqrs except for trust, which are the means and sds. Likert-scale questions are converted to scale on 1-5.