THE WHOLE IS OTHER THAN THE SUM OF ITS PARTS: SENSIBILITY ANALYSIS OF 360° URBAN IMAGE SPLITTING

Benjamin Beaucamp, Thomas Leduc, Vincent Tourre, Myriam Servières

THE WHOLE IS OTHER THAN THE SUM OF ITS PARTS: SENSIBILITY ANALYSIS OF 360° URBAN IMAGE SPLITTING

B. Beaucamp[1,*], T. Leduc[1], V. Tourre[1], M. Servières[1]

[1] Nantes Université, ENSA Nantes, École Centrale Nantes, CNRS, AAU-CRENAU, UMR 1563, F-44000 Nantes, France
{benjamin.beaucamp, vincent.tourre, myriam.servieres}@ec-nantes.fr, thomas.leduc@crenau.archi.fr

Commission IV, WG IV/3

KEY WORDS: Visual Urban Perception, 360° imagery, Deep Learning, Computer Vision

ABSTRACT:

360° imagery has been increasingly used to estimate the subjective qualities of the urban space, such as the feeling of safety or the liveliness of a place. These spherical panoramas offer an immersive view of the urban scene, close to the experience of a pedestrian. In recent years, Deep Learning approaches have been developed for this estimation task, only using flat images because these images are easier to annotate and process with standard CNNs. Thus to qualify the whole urban space, the panoramic images are divided into four flat sub-images that can be processed by the trained neural networks. The sub-images cover the 360° field of view, e.g. front, back, left, and right views. The four scores obtained are averaged to represent the level of the quality at the location of the panorama. However, this split introduces a bias since some elements of the urban space are halved over two images and the global context is lost. Based on the Place Pulse 2.0 dataset, this paper investigates the impact of splitting 360° panoramas on the perceptual scores predicted by neural networks. For each panorama, we predict the score for thirty-six overlapping sub-images. The scores were shown to have high variability and be highly dependent on the direction of the camera for the perspective images. This indicates that four images are not sufficient to capture the complexity of the perceptual qualities of the urban space.

1. INTRODUCTION

1.1 360° images to evaluate Visual Urban Perception

Visual urban perception aims at characterizing the urban space depending on the impact of its visible components on a pedestrian. Recently, thanks to the wide availability of imagery of the urban space at the pedestrian level, this problem has received more and more attention (Ibrahim et al., 2020, Biljecki and Ito, 2021). The use of images makes the assessment less cumbersome than the classical in-situ assessments, which require an expert to go on-site with a group of participants. Instead, this evaluation can be conducted in the lab with user experiments or with automatic tools. A major benefit of this approach is the ability to evaluate the urban space at a larger scale in a controlled environment that allows infinite replay, thanks to time-saving and tools that can automatically process a large number of images.

Platforms like Google Street View (GSV) or Baidu Street View enable the user, expert or not, to experience the urban space virtually. The users can look around thanks to 360° street view imagery and move spatially to explore a streetscape or a neighborhood. Some studies have demonstrated that to some extent, these services can be used as a replacement tool to evaluate the perception of the urban space (Rundle et al., 2011, Kelly et al., 2013). For instance, (Feng et al., 2021) showed that GSV can be used to assess the subjective perception of the urban environment, but is not reliable enough to assess the overall atmosphere.

One feature of these services that make this assessment possible is the presence of 360° images, which offer a panoramic and immersive view to the users who can see the surrounding urban space as a whole, compared to traditional pictures with a limited and arbitrarily oriented field of view (FOV). This is especially important since the subjective reaction to the urban space is not just a sum of perceptual responses to stimuli in the field of view of the pedestrian, but a global response to their surroundings (Feng et al., 2021). Therefore, the whole spherical view is important to assess the perception of the urban space since even turning around can heavily influence the user's reaction (Piga et al., 2021), the off-screen being a part of the scenery.

Such images have also been used in automatic tools that aim at assessing the perceptual qualities of the urban space, sometimes in conjunction with other data sources, for instance GIS (Yin, 2017). To process the images, the tools often rely on Deep Learning models, in two different ways: using semantic segmentation, also called *scene parsing*, to extract the content items of the image such as the amount of greenery, building, vehicles, etc. (Ma et al., 2021, Ramírez et al., 2021) or by directly training a neural network to predict perceptual qualities from an input image (Dubey et al., 2016, Wang et al., 2021, Liu et al., 2017a, Santani et al., 2018).

1.2 Splitting the panoramas for automatic processing

In the litterature, the current approach to process a panorama with a trained neural network is to split it into four sub-images that cover the whole horizontal FOV, because these networks are mainly designed to process flat images. For the rest of this paper, we call this method a 4-split. These four images are oriented in orthogonal directions, with a FOV of 90° that is close to the human FOV (see Figure 1), therefore the images don't overlap. The whole spherical image is not covered but most of the information in the observer's field of view is captured. Then, each image is scored by the neural network and the predictions are aggregated by averaging the scores, to reflect the perception

* Corresponding author

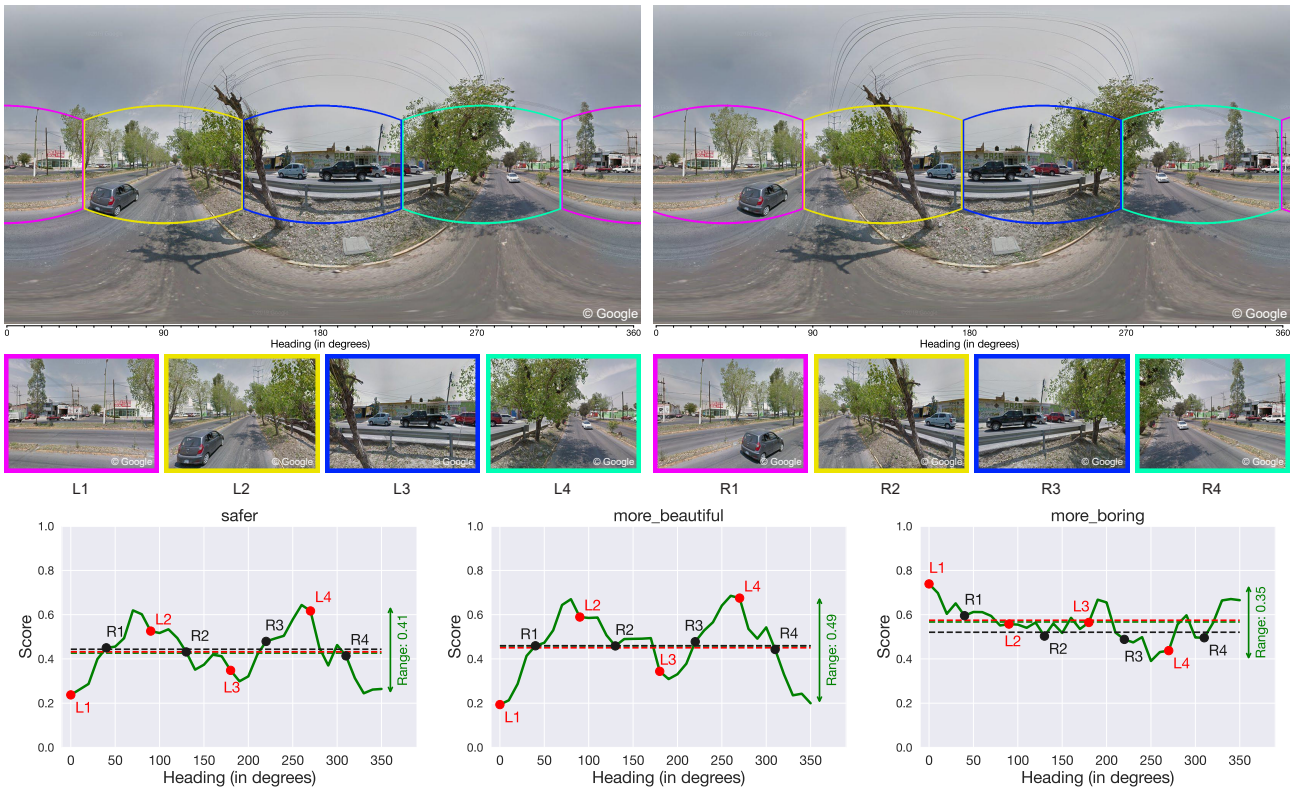


Figure 1. At the top left, a 360° panorama from Google Street View in equirectangular projection with four windows oriented at 0°, 90°, 180°, 270° (L1-L4) . At the top right, the same panorama is represented with four windows shifted by 40°(R1-R4), i.e. oriented at 40°, 130°, 220°, 310°. The perspective images corresponding to the windows are below each panorama. At the bottom, the scores for three of the six categories from the PP2 dataset are represented. The thirty-six scores obtained with the sliding window method are in green. The mean scores are represented by the horizontal dashed lines. In red is the mean of L1-L4, in black of R1-R4, and in green of the thirty-six scores. The mean score is about the same no matter the sampling chosen, but the scores have widely different distributions. L1-L4 windows give scores with high variability whereas the scores for R1-R4 are almost constant.

of the area. One major advantage of being able to process such a large amount of images automatically is the ability to qualify the urban space at the city scale. As a consequence, the scores need to be aggregated to be visualized at the city scale, using maps for instance. Two aggregation levels are commonly used: at the panorama-level (Blečić et al., 2018) (average score over the four sub-images from a panorama) or at the street-level (Ji et al., 2021, Yao et al., 2021, Zhang et al., 2018) (average score of all the perspective images in the street).

For models that rely on semantic segmentation, splitting a panoramic image into multiple perspective images is not an issue as the goal of this method is to obtain an accurate count of pixels for each category. However, the same remark cannot be easily extended to the tools that directly predict a perceptual quality from an input image of the urban space. As mentioned in (Feng et al., 2021, Piga et al., 2021), the panoramic images should be used as a whole to accurately predict individual reactions, which is not possible with the trained networks and datasets available. Thus, the only method currently available to characterize a place is to break it down into smaller pieces and analyze each one of them. But this approach is based on a very strong working assumption insofar as it ignores all the off-screen information for each piece.

1.3 Datasets for Visual Urban Perception

One of the most popular datasets for this task is Place Pulse 2.0 (PP2) (Dubey et al., 2016). It was created to train neural networks that are able to predict subjective and sensitive image qualities. More than 110,000 images from all over the world have been labeled by volunteers through a website. The labeling was achieved by comparing two images and indicating the one that best satisfies the given criteria. The six categories are *safer*, *livelier*, *more beautiful*, *more boring*, *more depressing* and *wealthier*. Although this dataset is very rich and one of the very few available, it has one major drawback with respect to what has been presented previously. The images used to qualify the spaces are traditional flat images with limited FOV, not 360° images. As a result, during their evaluation, the users had no idea of the context of the image to be qualified. They qualified only an image of a place, not a set of images nor a physical place. Other datasets have been created with a similar methodology but at a much smaller scale, with a few thousand images annotated to focus on a single city (Acosta and Camargo, 2019, Yao et al., 2021, Quercia et al., 2014, Candeia et al., 2017). However, it is understandable that due to technical reasons, these datasets are designed in this way as usually the CNNs do not take a full spherical image as input. Rather, they use *perspective images*, which are images with a limited field of view, as captured by traditional monocular cameras. The use of perspective images can be explained by several reasons: 1) the majority of neural networks in computer vision are designed to use perspective images, 2) it is much easier to collect data by asking participants to compare two perspective images rather than two spherical images, and 3) it is possible to split a panorama in multiple perspective images so we should be able to analyze panoramic images

with such a neural network. Therefore it could explain why, to our knowledge, there is no existing dataset with tagged 360° images so far.

1.4 Aim of this study

The use of 360° images to predict perceptual qualities of the urban space is a necessity as they contain much more information than a simple perspective image (Piga et al., 2021). They also offer an immersive point of view that can simulate the experience of a pedestrian. However, due to the difficulty of annotating spherical images, current Machine Learning approaches do not use the full 360° images as input. Instead, the panorama is split into sub-images that are processed independently, but previous works have paid little attention to the way of performing this split. In this paper, we perform a sensibility analysis of 360° image splitting and study how it influences the ranking of places per perceptual quality. Our main contributions are the following:

- An extension of the PP2 dataset to fetch and predict on the 360° images corresponding to the original "flat" ones (the code and trained models are available).
- A sliding window method to obtain a fine distribution of the scores over a panorama and to analyse the impact of the splitting.

We report two main findings:

- The scores at a panorama-level have high variability.
- When a panoramic image is split into four sub-images, the distribution of the four scores is highly dependent on the main direction of the points of view.

These results question the use of the panorama splitting method in a naive way and the use of the mean as an aggregation measure to synthesize the score of a perceptual quality at a given location.

2. METHODS

2.1 Dataset – Place Pulse 2 (PP2)

The PP2 dataset was created by (Dubey et al., 2016) to train neural networks to estimate the perceptual qualities of the urban space. 110,988 locations were chosen in fifty-six cities from twenty-eight countries across the world, making this dataset the most comprehensive one available yet. At each location, one perspective image was downloaded from the GSV API, thus giving an oriented (and partial) view of the surroundings. The orientation of the camera was not controlled by researchers during the data collection, so the API chose one that was directed towards the requested location, identified by its latitude and longitude. Then, annotations were crowdsourced online from 80,000 volunteers, from 2013 to 2016. Participants were presented two images and asked one question from six possible ones: "Which place looks safer?", "Which place looks livelier?", "Which place looks more boring?", "Which place looks wealthier?", "Which place looks more depressing?", and "Which place looks more beautiful?". To answer the question, the participant could choose either the left image, right image, or none if they were deemed of equal quality.

2.2 From PP2 perspective images to GSV panoramas

Each image in PP2 has a size of 400×300 pixels with a FOV of 90° and corresponds to a given position (informed by its geographical coordinates, in this case, its latitude and longitude). These perspective images were collected via the GSV API between 2007 and 2012. To perform our study, we "extended" the PP2 dataset: we chose the same locations as the ones in PP2, but we fetched the full 360° panorama at each location, instead of just one image with a limited FOV.
A 360° image is a sphere centered around the camera, and there are many ways to flatten this spherical image. The GSV API uses the equirectangular projection (see Figure 1). Given a camera orientation, a perspective image can be computed from the equirectangular panorama. Ever since its launch in 2007, Google collected panoramas multiple times at the same locations which highlights the evolution of the place through time. Depending on the location, there is an update of images each year or every few years. The GSV API provides metadata for each panorama including its precise geographical location, the date at which the image was taken, and a unique ID.
To retrieve the panorama from which comes each image in PP2, we need to know the unique ID of the panorama or its date. However, the date corresponding to each image in the PP2 dataset is not available, nor its unique ID, which means that we cannot easily fetch the matching panorama and we cannot, therefore, reproduce the same process.
To overcome this limitation in the dataset, we first collected all the panoramas available between 2007 and 2014 at each of the 110,998 PP2 image locations using the GSV API, resulting in a total of 408,886 panoramas. We then used feature matching with ORB descriptors (Rublee et al., 2011) between the panoramas and the PP2 images to find the correct or at least the closest panorama for each image. We were not able to find a matching panorama for each image in the PP2 dataset, as the GSV API sometimes did not return any data, or the feature matching step did not yield a good match. We were able to identify 95% of the sites, or, more precisely, we were able to find a matching 360° panorama for 95% of the perspective images in the PP2 dataset.
The zoom parameter in the API allows the user to choose from multiple resolutions, ranging from 416×208 pixels (zoom=0) to 13,312×6,656 pixels (zoom=5) for imagery before 2017. As our network is trained with 400x300 images with a FOV of about 90°, we downloaded panoramas of size 1,664×832 pixels (zoom=2), which proved to be enough to obtain perspective images of sufficient quality for inference.

2.3 Ranking Streetscore-CNN (RSS-CNN)

We used the RSS-CNN architecture introduced by (Dubey et al., 2016) to score the images. For an input perspective image, the network predicts a score for one of the six aforementioned perceptual qualities: *safer*, *livelier*, *more beautiful*, *more boring*, *more depressing* and *wealthier*.
The network is composed of two parts: Streetscore-CNN (SS-CNN) and a ranking sub-network. SS-CNN takes as input a pair of images, extracts their features, fuses them and determines the winning image for a given perceptual quality. The ranking sub-network is an addition to SS-CNN. It uses the extracted features to give a score for each image. The ranking network is useful as it gives an absolute score for an input image, so we do not need to rely on pairwise comparisons. The two parts are trained jointly on the PP2 dataset, using the pairwise comparisons collected via the crowdsourcing platform. Once the network is trained, we only rely on the ranking part to score the images.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-33-2022 | © Author(s) 2022. CC BY 4.0 License.

We trained our own RSS-CNN networks following the procedure described in (Dubey et al., 2016), as the code and model weights were not released. Next paragraph presents the procedure and the numerical values that we used for the parameters. They may differ from the original procedure as it has never been fully described in a publication.
For each perceptual quality, the pairs rated as "equal" by the participants were first discarded, then the dataset was split in the same proportions, 65% for training, 5% for validation and 30% for testing. We used VGG19 (Simonyan and Zisserman, 2014) pre-trained on ImageNet (Deng et al., 2010) for the feature extractor as it was reported by the authors to have the best performance. We trained the network with a batch size of 32 for 2 epochs, with a learning rate of 0.0001. The lambda parameter in the ranking loss was set to 1. A network was trained for each perceptual quality, with the same hyperparameters.

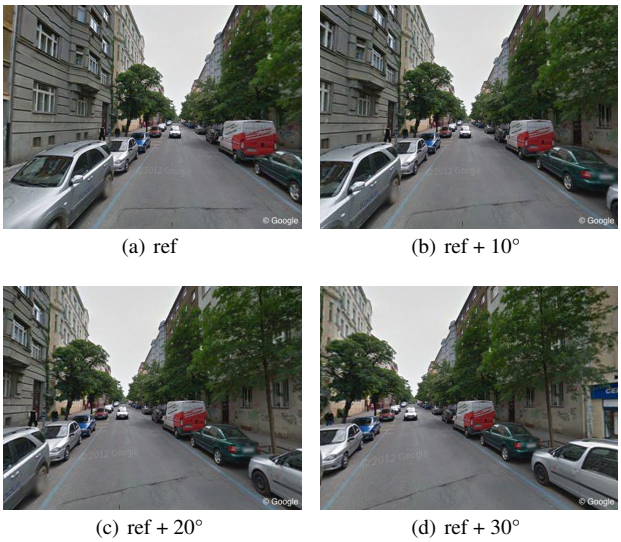
The performance of the network is evaluated with a standard accuracy measure since the task can be seen as a binary classification task. Our networks achieved an accuracy of 67.0% for *safer*, 66.0% for *livelier*, 62.2% for *more boring*, 68.7% for *wealthier*, 66.3% for *more depressing* and 69.8% for *more beautiful*. The accuracy scores are slightly lower than the ones reported by (Dubey et al., 2016), who achieved an accuracy of 73.5% for *safer*, 70.3% for *livelier*, 66.1% for *more boring*, 65.7% for *wealthier*, 62.8% for *more depressing* and 70.2% for *more beautiful*. This could be explained by the small differences in the training procedure or a different training/validation/testing split. However, the scores are similar to the ones obtained by other researchers who also trained their own RSS-CNN networks (Zhang et al., 2021, Xu et al., 2019). The code, trained model weights and scores on each panorama are available[1].

2.4 Sliding window method

To study the distribution of the scores on the panoramas, we chose a "sliding window" method. A perspective image is a "normal view" of a scene, as captured by regular monocular cameras. It can be seen as a "window" on the 360° sphere, i.e. a limited or partial view of the surroundings, as seen in Figure 1). This window is defined by three parameters: the heading (rotation from left to right), pitch (rotation from up to bottom), and FOV. All the perspective images had the following parameters: a FOV of 90°, a pitch angle of 0° (similar to a pedestrian looking straight ahead), with a resolution of 400×300. We slid this window on the whole panorama by changing the heading with a fixed step, and for each image, we used the RSS-CNN networks to obtain the six perceptual scores on this image.
We chose a heading step of 10° for several reasons: it is much lower than the 90° angle that is often used, allowing for a large overlap between two adjacent images (see Figure 2), it gives enough data points to compute statistical indicators, and the time necessary to perform inference with the six networks on the more than 100,000 panoramas was acceptable (ten days were required using an Nvidia Quadro P2200 graphics card). The scores obtained with the classical 4-split can be seen as a subsampling of the scores that we obtained with the sliding window method.

We pre-processed the data by removing outliers, identified as the scores that fell outside the 99.9% range. We obtained a total of 22,432,176 scores across all six qualities. For each quality, we normalized the range of the scores to be $[0, 1]$, 1 being the highest possible score for a given quality, and 0 being the lowest possible score.

[1] https://doi.org/10.5281/zenodo.6409860



(a) ref (b) ref + 10° (c) ref + 20° (d) ref + 30°

Figure 2. Four consecutive images with the sliding window method, obtained by repeatedly offsetting the camera heading by an angular step of 10° on the GSV panorama.

3. RESULTS

In this section, we examine the distributions of the scores predicted by the RSS-CNN networks on the 100k+ panoramas. These scores are in $[0, 1]$ due to normalization, and each panorama has a total of thirty-six scores for each quality, one for each heading in $[0°, 10°, \ldots, 350°]$. First, we take a look at the distribution of the thirty-six scores as a whole to show how much a perceptual quality varies on a 360° view. Then, we come back to the classical 4-split and demonstrate how the orientation of the four views impacts the distribution of the predicted four scores.

3.1 Distribution of the scores on a panorama

To summarize the scores predicted on the images resulting from the panoramic splitting, previous works usually compute the average of these scores (Blečić et al., 2018, Ji et al., 2021, Yao et al., 2021, Zhang et al., 2018). Often, this choice is made with little justification, which led us to wonder whether it is a good indicator of the perceptual quality at the location of the panorama. We chose to study how much the scores were spread on a 360° panorama. To do so, we computed the range of the thirty-six scores on each panorama. The range is defined as the peak to peak difference (i.e. the difference between the maximum score and the minimum score) on a panorama, thus indicating how far away the extremal scores are from each other. Figure 3 shows the distribution of these ranges over the whole dataset. The distributions are similar regardless of the quality chosen.

The scores were normalized between 0 and 1 so the maximum theoretical range is 1, which would be achieved if on the same panorama one image had a perfect score of 1 and another one a score of 0. The median is around 0.3 for each quality, meaning that for half of the panoramas the difference between the highest score and the lowest score is almost equal to a third of the maximum range. These results indicate that scores are

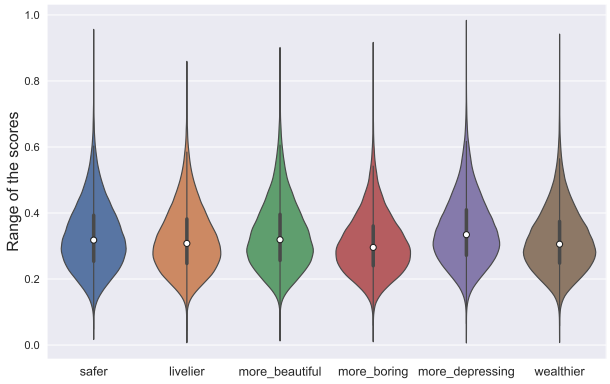


Figure 3. Distributions of the range of the thirty-six scores of each panorama, for each quality.

widely dispersed across panoramas, questioning the use of the mean as a relevant indicator to summarize scores across a panorama.

3.2 Impact of the panorama cut-out

A 4-split is a set of four perspective images extracted from a spherical 360° view. If we call d the direction of the camera for one of the images, the direction of the camera for the other ones are given by $d+90°$, $d+180°$, $d+270°$ (modulo 360°). Thus, a 4-split can be defined solely by the direction of one of the four images, called the principal direction for the rest of this section. To study the impact of the principal direction in a 4-split for a given panorama, we compared the standard deviation of the four scores between a *reference* 4-split and a *shifted* 4-split (see Figure 4). The standard deviation measures how much the values are spread around the mean, so we chose it as an indicator of the dispersion of the scores. For instance, on Figure 1, the *safety* scores for the reference 4-split are fluctuating between 0.2 and 0.6 (L1-L4), while those of the 40° shifted 4-split are almost constant to 0.4 (R1-R4).
More precisely, for a given shift s in degrees, we studied the impact of this shift on each panorama. To do so, we selected all pairs of 4-split (h_1, h_2) such that h_2 corresponds to h_1 shifted by $s°$. Then we computed the relative variation (RV_{std}) between the standard deviation of the reference 4-split scores and the scores of the shifted one, as given by:

$$RV_{std} = \frac{|\text{std}_{\text{ref}} - \text{std}_{\text{shifted}}|}{\text{std}_{\text{ref}}}$$

where std_{ref} (resp. $\text{std}_{\text{shifted}}$) is the standard deviation of the four scores on the images with the reference (resp. shifted) 4-split.

Finally, for each panorama, we took the median of these relative variations as the measure of the impact of the shift on this panorama. Here, the median is more informative than the mean since these relative variations can take large values in cases like the one shown in Figure 1. The distribution of these median scores for different shifts are represented on Figure 5, for the *safer* attribute. On this plot, we removed the data points outside 1.5 interquartile range for clarity, as they are commonly recognized as being outliers.

The median ranges from 20% to 35% for the studied shifts and the shift has more impact as its value is closer to 45°. It was predictable as that is the point where the four images with the

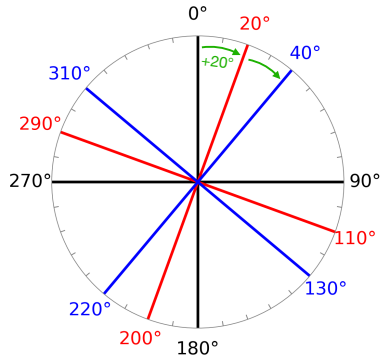


Figure 4. Top view of a panorama with the ticks representing the heading azimuths. Lines of the same color represent the four camera directions for the four perspective images in a 4-split. The red 4-split corresponds to the black one shifted by 20°, and the blue one is equal to the red one shifted by 20°.

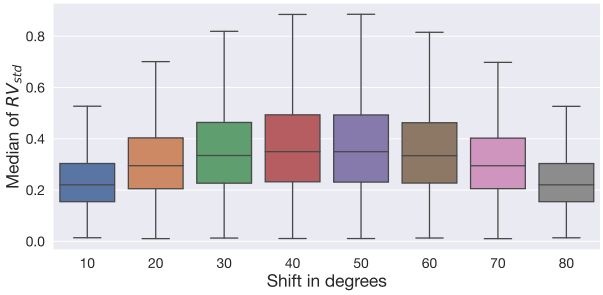


Figure 5. Impact of a shift of camera direction on the standard deviation of the scores when using only a 4-split, for the *safety* attribute.

shifted headings are the most different than the reference images. Still, even with a small shift, the difference in standard deviation is high. The same conclusions can be made for the five other qualities (see Table 1).

	median of RV_{std}		median of RV_{mean}	
	mean	std	mean	std
safer	0.320	0.160	0.042	0.026
livelier	0.317	0.162	0.041	0.024
wealthier	0.321	0.160	0.041	0.025
more beautiful	0.316	0.159	0.044	0.044
more boring	0.326	0.160	0.044	0.028
more depressing	0.324	0.159	0.053	0.032

Table 1. Aggregated values of the median of RV_{std} and RV_{mean} for each perceptual quality.

This measure is conservative since we study the impact of the shift over all possible reference four headings and use the median to report the impact on a panorama. In practice, a shift can have a huge impact on the distribution of the four scores as can be observed in Figure 1.

On the contrary, an interesting result is that when a 4-split is used to analyze a panorama, the mean of the four scores remains relatively stable when a shift is applied. By using the same measure of relative variation, this time on the mean (RV_{mean}), we see on Figure 6 that the mean varies by about 4% (median score), and no more than 10%, for any shift. This relative stability is shared by all six qualities (see Table 1).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-4-2022-33-2022 | © Author(s) 2022. CC BY 4.0 License.

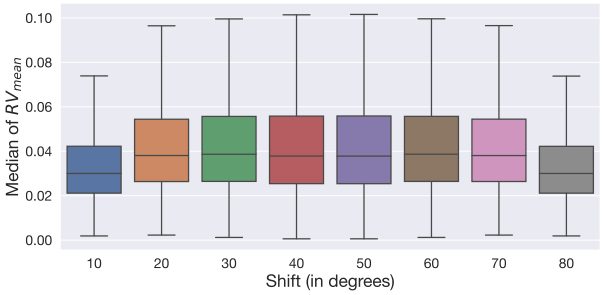


Figure 6. Impact of a shift in camera direction on the mean score when using only a 4-split, for the *safety* attribute.

4. DISCUSSION

4.1 Findings

First, the analysis of the predicted scores demonstrated that the scores on a panorama (i.e. for one geographical location) are significantly spread out. Previous studies have often used the mean of these scores to reflect how a quality is perceived at a location (Blečić et al., 2018) or in a street (Ji et al., 2021, Yao et al., 2021, Zhang et al., 2018), but our results show that this statistic is not sufficient per se as a single indicator. Extremums play an important part as they report on very positive or very negative areas of the surrounding space. These areas play an important part in the overall feeling of the pedestrian, so they should be accounted for, which is why the mean score alone is not enough. Moreover, a metric that characterizes the dispersion of the scores can be useful as it informs on whether a place is contrasted or not. The six perceptual qualities in this dataset are used to give an overall idea of the ambiance of a place, but just by focusing on a single quality, we showed that they cannot be simply evaluated by averaging the scores on a few images as their distributions are complex.
Second, because of the variability of the scores, splitting the panorama into four can give drastically different results depending on the principal direction of the four images. We may explain this by the presence of objects and their location in the panorama. Several studies have shown that the presence of certain "objects" such as trees, cars, or pedestrians heavily influences the perceptual score of the image (Ramírez et al., 2021, Rossetti et al., 2019, Ji et al., 2021). Thus, if the object is split in half it will likely not have the same impact on the output of the network. The annotated data may also be subject to the same issue.

4.2 Study limitations

This work includes a few limitations. We used a single architecture for the neural network and did not test another one. However, it is a common architecture that other studies have used either to predict scores or as a baseline to compare to.
The predicted scores were also taken "as is", even though the trained networks, as any ML models, are known to have limited accuracy. To mitigate this limitation, we mainly compared the predicted scores between each other (as the network is trained to do, i.e. discriminate between two images) rather than analyzing the absolute scores.
We chose to focus our study on the PP2 dataset as it covers a wide range of urban streetscapes and is large enough to train robust neural networks. It is also the most commonly used dataset to study visual urban perception, since it does not focus on a particular city. As a consequence, our study suffers from the same limitations as PP2, but our results can still be used by other researchers interested in panoramic image splitting. For instance, the FOV and pitch are constant for all the images in the dataset. To ensure the validity of the scores at inference time with the trained neural networks, we made sure to only use images similar to the ones in the training set, i.e. images with the same size, FOV and pitch. Because of this, even if multiple images are used on the horizontal axis, the top and bottom parts of the sphere are not taken into account. It would be interesting to explore the relationship between the FOV and the predicted perceptual quality, but there is no guarantee on the validity of the scores predicted with an image with a FOV smaller or greater than the one used in the training set, so this would require a new dataset.
Finally, we only analyzed the distribution of the scores on a panorama, and not at a larger spatial scale, although some works aggregate the scores at the street-level (Ji et al., 2021, Yao et al., 2021, Zhang et al., 2018). Nevertheless, our conclusion regarding the range of the scores should still hold, i.e. the distribution of the scores in a street should present a large difference between the extremum.

5. FURTHER WORKS AND CONCLUSION

5.1 Usage of panorama splitting with current datasets

With the current dataset and neural network architectures available, the most sensible approach to process a 360° image is still to split it into multiple images. In the future, care should be taken on how a panorama is subdivided into sub-images as it can heavily impact the estimation of the distribution of the scores over the 360° view. In the literature, the 4-split is often chosen arbitrarily with the cardinal directions (true north, west, south, east). Further studies are needed to find a more meaningful split. For instance, a split based on the morphology of the urban space could solve this issue, by using viewpoints aligned with the street (facing front, back, right, and left) or the main openings. Another idea would be to split the panorama based on the content of the image, which can be estimated via Semantic Segmentation (Garcia-Garcia et al., 2018, Ye et al., 2019, Ramírez et al., 2021). We observed that a subsampling of the perceptual scores with only four points is not reliable enough to estimate the distribution of the perceptual scores on a panorama. Therefore, researchers should consider using more than four images to better represent the complexity and the richness of the perceptual scores and be less subject to viewpoint dependence.
We have shown that the scores on a panorama have a complex and rich distribution, so further work is planned to explore how the scores can be aggregated to reflect the overall quality of a place. This aggregation step is necessary to be able to report the results efficiently to urban planners and collectivities at the city scale, and more widely to all the stakeholders of urban design. For instance, the level of each quality can be represented on a map, and areas with poor scores can be identified with spatial clustering analysis to better inform future urban projects (Yao et al., 2021).

5.2 Towards the use of full 360° images

5.2.1 Using the whole panorama For now, most datasets with subjective annotation contain only flat images with a FOV of 90°. As a consequence, neural networks trained with these datasets are also limited and can only process images with this FOV. It would be interesting to use images with a larger FOV

to encompass more information in a single picture, as the ambiance of a place has been shown to be a response to the overall surrounding area. This would require a new dataset to train the networks, but a similar online crowdsourcing platform could be built to ask participants to compare images with a FOV of 120°, as they can still be considered to be "flat".
Otherwise, using images with a higher pitch (looking up) should be studied. The images would often show the top of buildings and the sky, which are important as they can be used to compute the sky view factor, an estimator of the openness (or the opposite, enclosure) of a place (Ma et al., 2021).
Full 360° images could also be used to avoid splitting the panorama, but it would be much more labor-intensive to collect such a dataset and train a network. A user would need to be able to explore the two 360° images to make a choice, which would take longer than just comparing two perspective images and introduce methodological issues due to the unintuitive comparison of two 360° images.

5.2.2 Dependency between the viewpoints Instead of having a label for one image or a label for a pair of images, (Liu et al., 2017b) proposed to use a dataset with labels associated with *places*. Here, the label represents the level of a perceptual quality (e.g. safety) for the physical place, represented by a group of images. These images represent different views of the place, in this case eight overlapping perspective images extracted from an equirectangular panorama. Such labels were derived from crime records from several cities in the United States. With similar data, (Suel et al., 2019) presented a neural network that processes four images of a place at the same time (a 4-split of an equirectangular panorama) to predict the label associated with the place. These methods are an in-between as they can process more than one perspective image, but not a full 360° panorama. Still, they allow for the score at a geographical location to be predicted directly from different views of the urban space, instead of having to aggregate multiple scores predicted on perspective images.
Also, it would be interesting to process the sub-images together to account for the spatial dependency. A pedestrian does not experience the urban space through a collection of independent viewpoints. For instance, after seeing the image at a heading of 0°, the image at 90° corresponds to the user moving its head to the right. There is a continuity between each viewpoint and they are interdependent. Even more, the spatial proximity between two panoramas can be accounted for to represent a pedestrian following a path.

5.3 Explaining the scores with the sliding window technique

A few studies have tried to explain the relationship between the content of the image and the perceptual score predicted by the neural network (Quercia et al., 2014, Porzi et al., 2015). Further work should explore the use of this sliding window technique to better understand this relationship. When a large variation in score is identified for two consecutive windows, it could indicate that the elements that appeared in the frame or disappeared from the frame are correlated with the perceptual quality. A better understanding of the relationship between objects and perceptual qualities will also help to find a better way to split panoramic images.

5.4 Conclusion

Many attempts at using 360° imagery for the prediction of perceptual qualities of the urban space chose to split panoramas into several "flat" sub-images with a limited FOV. Researchers carefully split the panoramas into four images that cover the whole surrounding area (Candeia et al., 2017, De Nadai et al., 2016). However, there is little work studying the validity of such a split and the subsequent aggregation of the scores predicted on each image. Our goal was to study the distribution of the scores on a 360° image and the impact of splitting a panorama on the estimation of a perceptual score at a location. To do so, we slid a window with a FOV of 90° along the full panorama with a step of 10° and used an RSS-CNN network to predict a score on each window. Our study showed that the perceptual qualities are too complex to be summarized with a mean score, and that the main direction of the four split heavily influences the scores obtained.

ACKNOWLEDGEMENTS

The authors thank the École Centrale de Nantes and the Région Pays de la Loire through the Atlanstic 2020 programme for the financial support of this project. This work was performed by using HPC resources of the Centrale Nantes Supercomputing Centre on the cluster Liger and supported by a grant from the Institut de Calcul Intensif (ICI) under the project ID OG2102040 / 2021.

REFERENCES

Acosta, S., Camargo, J. E., 2019. Predicting city safety perception based on visual image content. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11401 LNCS, 177–185. http://arxiv.org/abs/1902.06871.

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

Blečić, I., Cecchini, A., Trunfio, G. A., 2018. Towards automatic assessment of perceived walkability. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10962 LNCS, Springer Verlag, 351–365.

Candeia, D., Figueiredo, F., Andrade, N., Quercia, D., 2017. Multiple images of the city: Unveiling group-specific urban perceptions through a crowdsourcing game. *HT 2017 - Proceedings of the 28th ACM Conference on Hypertext and Social Media*, Association for Computing Machinery, Inc, New York, NY, USA, 135–144.

De Nadai, M., Vieriu, R. L., Zen, G., Dragicevic, S., Naik, N., Caraviello, M., Hidalgo, C. A., Sebe, N., Lepri, B., 2016. Are Safer Looking Neighborhoods More Lively? A Multimodal Investigation into Urban Life. *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, 1127–1135. http://arxiv.org/abs/1608.00462.

Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, Li Fei-Fei, 2010. ImageNet: A large-scale hierarchical image database. 248–255.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C. A., 2016. Deep learning the city: Quantifying urban perception at a global scale. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905 LNCS, Springer Verlag, 196–212.

Feng, G., Zou, G., Piga, B. E. A., Hu, H., 2021. The Validity of Street View Service Applied to Ambiance Perception of Street: A Comparison of Assessment in Real Site and Baidu Street View. *International Conference on Applied Human Factors and Ergonomics*, 740–748. https://link.springer.com/chapter/10.1007/978-3-030-80829-7_91.

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., Martinez-Gonzalez, P., Garcia-Rodriguez, J., 2018. A survey on deep learning techniques for image and video semantic segmentation.

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

Ji, H., Qing, L., Han, L., Wang, Z., Cheng, Y., Peng, Y., Brovelli, M. A., Kainz, W., 2021. A New Data-Enabled Intelligence Framework for Evaluating Urban Space Perception. *ISPRS International Journal of Geo-Information 2021, Vol. 10, Page 400*, 10(6), 400. https://doi.org/10.3390/ijgi10060400.

Kelly, C. M., Wilson, J. S., Baker, E. A., Miller, D. K., Schootman, M., 2013. Using Google Street View to audit the built environment: Inter-rater reliability results. *Annals of Behavioral Medicine*, 45(SUPPL.1), 108. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3549312/.

Liu, L., Silva, E. A., Wu, C., Wang, H., 2017a. A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems*, 65, 113–125.

Liu, X., Chen, Q., Zhu, L., Xu, Y., Lin, L., 2017b. Place-centric visual urban perception with deep multi-instance regression. *MM 2017 - Proceedings of the 2017 ACM Multimedia Conference*, Association for Computing Machinery, Inc, 19–27.

Ma, X., Ma, C., Wu, C., Xi, Y., Yang, R., Peng, N., Zhang, C., Ren, F., 2021. Measuring human perceptions of streetscapes to better inform urban renewal: A perspective of scene semantic parsing. *Cities*, 110, 103086.

Piga, B., Stancato, G., Boffi, M., Rainisio, N., 2021. Emotional clustered isovist. Representing the subjective urban experience Introduction. D. D. Mascio (ed.), *EAEA15 ENVISIONING ARCHITECTURAL NARRATIVES*, the University of Huddersfield, Huddersfield,, 163–173.

Porzi, L., Buló, S. R., Lepri, B., Ricci, E., 2015. Predicting and understanding Urban perception with convolutional neural networks. *MM 2015 - Proceedings of the 2015 ACM Multimedia Conference*, Association for Computing Machinery, Inc, New York, NY, USA, 139–148.

Quercia, D., O'Hare, N., Cramer, H., 2014. Aesthetic capital: What makes london look beautiful, quiet, and happy? *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, Association for Computing Machinery, New York, NY, USA, 945–955.

Ramírez, T., Hurtubia, R., Lobel, H., Rossetti, T., 2021. Measuring heterogeneous perception of urban space with massive data and machine learning: An application to safety. *Landscape and Urban Planning*, 208, 104002.

Rossetti, T., Lobel, H., Rocco, V., Hurtubia, R., 2019. Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. *Landscape and Urban Planning*, 181, 169–178.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB: An efficient alternative to SIFT or SURF. *Proceedings of the IEEE International Conference on Computer Vision*, 2564–2571.

Rundle, A. G., Bader, M. D., Richards, C. A., Neckerman, K. M., Teitler, J. O., 2011. Using google street view to audit neighborhood environments. *American Journal of Preventive Medicine*, 40(1), 94–100. https://pubmed.ncbi.nlm.nih.gov/21146773/.

Santani, D., Ruiz-Correa, S., Gatica-Perez, D., 2018. Looking South: Learning Urban Perception in Developing Cities. *ACM Transactions on Social Computing*, 1(3), 1–23. https://dl.acm.org/doi/10.1145/3224182.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://arxiv.org/abs/1409.1556v6.

Suel, E., Polak, J. W., Bennett, J. E., Ezzati, M., 2019. Measuring social, environmental and health inequalities using deep learning and street imagery. *Scientific Reports 2019 9:1*, 9(1), 1–10. https://www.nature.com/articles/s41598-019-42036-w.

Wang, R., Ren, S., Zhang, J., Yao, Y., Wang, Y., Guan, Q., 2021. A comparison of two deep-learning-based urban perception models: which one is better? *Computational Urban Science*, 1(1), 1–13. https://doi.org/10.1007/s43762-021-00003-0.

Xu, Y., Yang, Q., Cui, C., Shi, C., Song, G., Han, X., Yin, Y., 2019. Visual urban perception with deep semantic-aware network. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11296 LNCS, Springer Verlag, 28–40.

Yao, Y., Wang, J., Hong, Y., Qian, C., Guan, Q., Liang, X., Dai, L., Zhang, J., 2021. Discovering the homogeneous geographic domain of human perceptions from street view images. *Landscape and Urban Planning*, 212, 104125.

Ye, Y., Zeng, W., Shen, Q., Zhang, X., Lu, Y., 2019. The visual quality of streets: A human-centred continuous measurement based on machine learning algorithms and street view images. *Environment and Planning B: Urban Analytics and City Science*, 46(8), 1439–1457.

Yin, L., 2017. Street level urban design qualities for walkability: Combining 2D and 3D GIS measures. *Computers, Environment and Urban Systems*, 64, 288–296.

Zhang, C., Wu, T., Zhang, Y., Zhao, B., Wang, T., Cui, C., Yin, Y., 2021. Deep semantic-aware network for zero-shot visual urban perception. *International Journal of Machine Learning and Cybernetics 2021*, 1–15. https://link.springer.com/article/10.1007/s13042-021-01401-w.

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., Ratti, C., 2018. Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180, 148–160.