Computational Quantitative Aesthetics Evaluation

Evaluating architectural images using computer vision, machine learning and social media

Victor Sardenberg[1], Mirco Becker[2]
[1,2]Leibniz Universität Hannover
[12]{sardenberg|becker}@iat.uni-hannover.de

This paper correlates two methods of aesthetic evaluation of architectural images utilising computer vision (CV) and machine learning (ML) for automating aesthetic evaluation: Calibrated aesthetic measure (CalAM) and aesthetic scoring model (ASM). From a database of images of proposals for a single location, users are invited to like or dislike it on social media to feed an ML model and calibrate an aesthetic measure formula (AMF). A possible application is to assist designers in making decisions according to the hedonic response given by users previously, enabling a faster way of popular participation.

Keywords: *Quantitative Aesthetics, Crowdsourcing, Aesthetic Measure, Computer Vision, Machine learning, Social Media.*

INTRODUCTION

This paper correlates two methods of aesthetic evaluation using computer vision (CV): (1) A calibrated aesthetic measure (CalAM) formula and (2) an aesthetic scoring model (ASM).

CalAM applies CV to segment architectural parts in images to quantify their relations and calculate an aesthetic measure. This method upgrades the previous work by George David Birkhoff (1933), Max Bense (1965) and Sigfried Maser (1968) to the contemporary computational capabilities.

ASM utilises CalAM and other quantified results from our CV analysis to train artificial neural networks (ANN) to predict aesthetic scores.

We crowdsourced an audience on social media to register their hedonic response to calculate an average score for each image and output an average hedonic score (AHS) for each image. That score was used as the ground truth reference to calibrate CalAM and train ASM. Online crowdsourcing constitutes now a compelling way of gathering empirical datasets fast and cheaply.

We utilise the images submitted to the Young Architects Program from MoMA PS1 as an image database for evaluation. It contains images of built and unbuilt projects by various architects since 1998 for the exact location: the entrance patio of the PS1, New York, USA.

A possible application of these methods for automatised evaluation is to assist designers in making decisions incorporating more popular suggestions among the audience of people that will utilize their designs daily. CalAM and ASM enable designers to harness crowdsource intelligence to train artificial intelligence to design more inclusive and diversified environments. This research focuses only on the visual experience of architecture.

STATE OF ART

There is a large body of research on the main topics that this research builds on:

- Aesthetic measure,
- Machine learning training for aesthetic evaluations, and
- Crowdsourcing empirical aesthetics.

Aesthetic Measure

This term was coined by the US-American mathematician George David Birkhoff (1884 – 1944) in 1933 in his book titled "Aesthetic Measure". He devises mathematical formulas to judge aesthetic objects, namely polygonal forms, tilings and vases, with empirical studies he conducted. In this book, his project is "to bring the basic formal side of art within the purview of the simple mathematical formula defining aesthetic measure" (1933, p. viii).

He describes the aesthetic experience as

> "[…] compounded of three successive phases: (1) a preliminary effort of attention, which is necessary for the act of perception, and which increases in proportion to what we shall call the complexity (C) of the object; (2) the feeling of value or aesthetic measure (M) which rewards this effort; and finally (3) a realisation that the object is characterised by a certain harmony, symmetry, or order (O), more or less concealed, which seems necessary to the aesthetic effect" (1933, p.4)

Therefore, the following formula for the aesthetic measure is proposed by Birkhoff:

$$\text{Aesthetic Measure} = \frac{\text{Order}}{\text{Complexity}}$$

This basic formula for the aesthetic measure is determined by the density of order relations in the aesthetic object. "The definition of the beautiful as that which gives us the greatest number of ideas in the shortest space of time (formulated by Hemsterhuis in the eighteenth century) is of analogous nature." (1933, p. 4). If we admit the validity of it and this formula, the fundamental aesthetic problem is: "Within each class of aesthetic objects, to define the order O and the complexity C so that their radio M = O/C yields the aesthetic measure of any object of the class." (Birkhoff, 1933, p. 4)

Each kind of aesthetic object gives rise to aesthetic feelings, which is *sui generis*. Therefore, each type of aesthetic object requires a different adaptation of the aesthetic measure formula.

His formula was taken up by Max Bense (1910 – 90), which proposed the concept of Information Aesthetics (Bense, 1965). Considering that objects have objective aesthetic states, Bense argued for an objective evaluation of works. Sigfried Maser, who studied under Bense and submitted his doctoral thesis "Numerische Ästhetik" (Numerical Aesthetics) in 1967, developed the concepts of order and complexity in a more objective way than Birkhoff's original method. Manfred Kiemle, another student of Bense, applied this concept to architectural facades (Kiemle, 1967).

To quantify shapes, vases, facades and aesthetic objects in general, it was necessary to simplify and compress them to a very restricted set of characteristics to be verified by a human. With the recent developments in CV, it is possible to revisit these ideas by embracing all visual data of bitmaps and fully automating its evaluation.

Machine Learning training for aesthetic evaluations

Computer vision and machine learning have been combined in multiple studies to understand how users perceive urban environments.

Verma, Jana and Ramamritham (2018) have developed mobile apps to capture images of neighbourhoods and applied semantic segmentation as inputs for neural networks to judge surroundings. Salesses, Schechtner and Hidalgo (2013) have utilised geo-tagged images to create quantitative measures of urban perception and characterise the inequality of cities. Dubey et al (2016) combine crowdsourcing with neural networks to produce urban perception data globally. Seresinhe, Preis and Moat (2017) explore online data from a game with neural networks to judge if outdoor images are scenic and beautiful. All these methods utilise semantic segmentation for training neural networks.

Crowdsourcing empirical aesthetics

There are plenty of examples of digitally enabled hedonic judgement online. Tinder, Instagram, Facebook, and TikTok are social media platforms where users like and dislike other users' images according to their aesthetic attributes. They also expect to be evaluated similarly. Applications like that allow to smoothly translate a qualitative judgement of images into quantities of likes or dislikes. These quantities can generate an average hedonic score (AHS).

Sardenberg and Becker (2019) built an online platform mimicking Tinder swipe left and right interface to crowdsource architectural image aesthetic judgement and feed the average score as the fitness criterion for evolutionary algorithms to navigate solution spaces.

MoMA PS1 Young Architects Program as image dataset

The MoMA PS1 Young Architecture Program is a regular architectural invited competition. Since 1998, five proposals have been yearly submitted to the museum for installations on the patio of the MoMA PS1 in Queens, New York. All submissions are available on MoMA´s website.

Considering the extensive collection of proposals for the same singular site (87 designs) and the availability of this archive for assessment, we selected a maximum number of two human point-of-view images of each proposal. No distinction was made regarding technique, which included photographs of built projects, computer-generated imagery (CGI), collages, paintings, and pictures of models. That totalised 141 images.

It is essential to consider a significant difference in the number of humans, the percentage of sky pixels and the presence of natural elements, which are crucial elements of analysis for scoring images in the above-referenced studies that approach images from semantic segmentation techniques.

METHODS

We introduce two methods for aesthetic quantification: (1) CalAM and (2) ASM. Each method is comprised of all or most of these sub-methods: (A) MSER and parts-from-wholes, (B) aesthetic measure formula and calibration, (C) crowdsourced AHS and (D) ANN training. Refer to Figure 1 for an overall view of our methods and sub-methods.

Compared to the previous efforts on image judgements with ANN, our methods only focus on the compositional aspects of images, not being influenced by semantic elements like the kinds of objects portrayed. When architects are designing, their decisions are primarily compositional, like (a) how parts relate to wholes, (b) how refer inside to

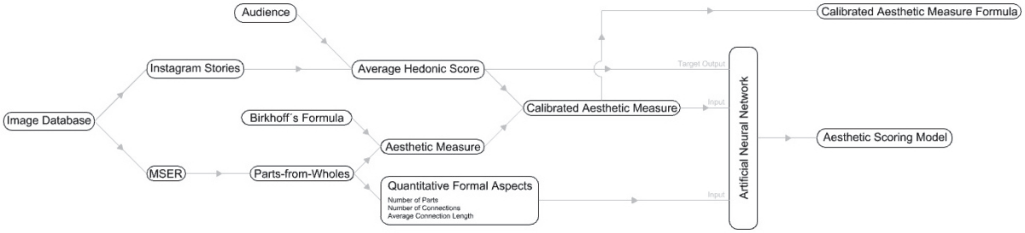


Figure 1
Diagram of methods and sub-methods described in this paper.

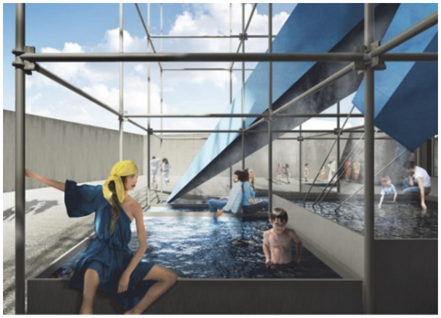


Figure 2
Wendy by HWKN

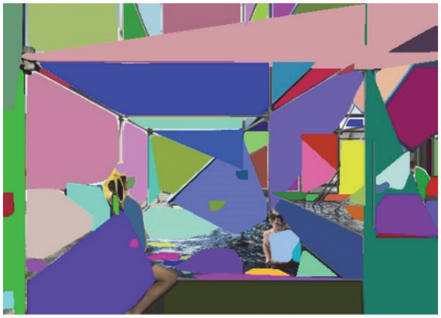


Figure 3
Coloured Regions

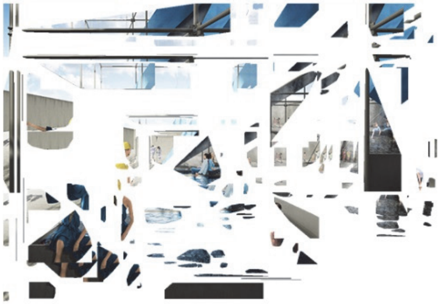


Figure 4
Diagram of Scaled Parts

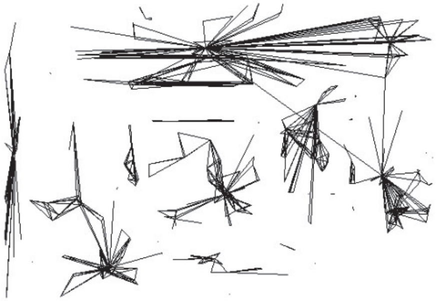


Figure 5
Diagram of Connectivity

“ outside, (c) how voids relate to masses, (d) figures and objects relate to the ground and (e) surfaces to volumes. The semantic elements of images considered in the previous methods are of little use during the design decision phase because architects are not defining, or indirectly, where crowds of people will be, the colour of the sky, the state of trees or the location of cars.

MSER and parts-from-wholes sub-method

We utilise MSER (Maximally stable extremal regions) as a CV algorithm because it recognises high-contrast regions in images, like characters on photos (Matas, 2004). We repurpose it to identify parts of architectural elements on images.

MSER converts multi-tone grayscale images into black or white binary pixels using multiple thresholds from entirely black to completely white. Therefore, it recognises regions, and when they are consistent among various thresholds, they are outputted. We calibrated such a system to coincide with architectural elements in photos.

We propose a sub-method, previously described on a poster at the Design Computing and Cognition 2022 Conference (in print), entitled parts-from-wholes. It translates the MSER’s regions into two kinds of diagrams with specific goals and outputs quantitative data about the image composition. To exemplify, we applied it here to an image of the HWKN´s winning proposal, “Wendy,” submitted for the 2012 edition.

The diagram of scaled parts (Figure 4) consists of the pixels included in the regions scaled. Each region is scaled by half resulting in autonomous parts sprawled through the graphic space. This is a powerful way of understanding the qualities of each part.

The connectivity diagram (Fig. 5) focuses on how the regions intersect. The algorithm checks if each region intersects with or includes all other regions completely. If this is the case, it draws a line from one centroid to the other one. Here it is only possible to see how the parts relate in the graphic space.

This sub-method extracts quantitative data from the image, as seen in table 1.

Table 1
The quantitative output of the parts-from-whole sub-method

Number of Parts	358
Parts Area Average	1261px
Minimum Part Area	68px
Maximum Part Area	25132px
Number of Connections	1222
Connection Length Average	72px
Maximum Length	388px

Crowdsourcing AHS sub-method

We propose a method of digitally recording the hedonic response of social media users. A hedonic response is a judgement of liking or disliking something (Shimamura, 2014).

Instagram Stories is utilised to reach a large audience in the current paper. Stories is a modality of sharing content where portrait-oriented slides are presented for 4 seconds and enable interactions like answering questions, participating in polls or liking the content. Each slide is only available for 24 hours.

One of the authors utilised his profile with 1873 followers to invite them to score the images using sliders from dislike to like. The experiment was shortly introduced, with a few slides explaining it. After the introduction, each slide contains one image and a slider where the user can rapidly answer how much they like or dislike its content. After 24 hours, the average scores are registered as an AHS that is used as the value accepted as ground truth for calibrating the aesthetic measure formula and the target output of the aesthetic scoring model. During 14 days, 10.716 scores of all 141 images were given by this audience.

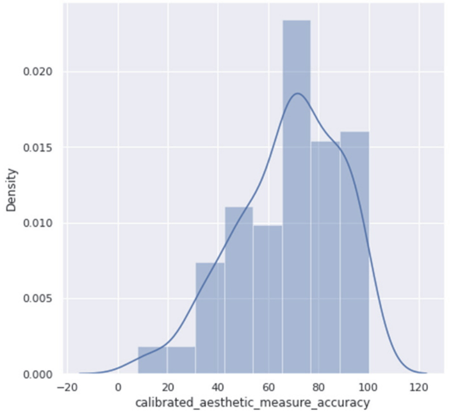
The audience is biased toward architects based mainly in Europe and South America in their 30s. We would argue that this bias is not necessarily problematic according to how the method here described is applied. We devise it being utilised to assist a designer in making decisions according to the average hedonic response of a particular audience (inhabitants of a specific neighbourhood where a new project will be built, for example). If all socio-cultural profiles of this audience are well represented, it doesn´t need to reflect a human inborn aesthetic evaluation.

Aesthetic Measure and Calibration sub-method

We adapted Birkhoff´s formula of aesthetic measure to score image compositions using the quantitative output of our parts-from-wholes sub-method as its terms.

Considering that "the aesthetic measure is determined by the density of order relations in the aesthetic object" (Birkhoff, 1933, p.4), we defined the number of connections and their length as indicators of order and the number of parts as an indicator of complexity, which results in the following formula:

$$\text{Aesthetic Measure} = \frac{\text{Order}}{\text{Complexity}} = \frac{\text{Connection Length Average} \times \text{Number of Connections}}{\text{Number of Parts} \times \sqrt{\text{Number of Pixels}}}$$



Graph 1
Histogram of Accuracy of the CalAM

To normalise the value across image resolutions, we divide by the square root of the number of pixels, that is, in the case of this paper, always 540 x 386px.

For Birkhoff, it is impossible to compare objects of different types. For example, according to him, "it is futile to compare a painting in oils with one in watercolours". Therefore, it is impossible to compare a photo of built projects with CGI of unbuilt ones. However, as Birkhoff maintained, “the two paintings might be compared, in respect to composition alone, by means of photographic reproduction”. This is what our method does: it compares the composition of different image-making techniques (Photos, CGI, collage…) employing bitmaps.

Birkhoff’s method requires a human to interpret the elements of order and complexity of an object, making it empirical, time-consuming and dependent on subjectivity. Sigfried Maser (1968) developed an application of this formula in a more objective way.

Our sub-method applies CV to remove human interpretation to respond to it. It recognises the parts of a composition, how they connect and calculates an aesthetic measure.

When the formula is applied to the compositional analysis of the images and compared to the AHS, the accuracy is only 38%. It is necessary to calibrate the weight of each term of the formula to approximate it to the AHS. We utilise the evolutionary solver of Galapagos to apply weights from 0 to 1 for each term of the formula. The calibrated formula presented ahead reached 66.98% accuracy:

$$\text{Aesthetic Measure} = \frac{0.\,82\ \text{Connection Length Average} \\ \text{x}\ 0.96\ \text{Number of Connections}}{0.\,29\ \text{Number of Parts} \\ \text{x}\ \sqrt{}\ \text{Number of Pixels}}$$

ANN training sub-method

We utilise supervised machine learning running to train an ANN using a sequential model. The structure of our neural network is described in figure 6.

The data inserted contains the connection length average, number of connections, number of parts, aesthetic measure and CalAM as inputs and AHS as target output. The correlation matrix is shown in table 2.

The data was split as 75% for the training set and 25% for the testing set, and the model was trained for 3000 epochs using Geoffrey Hinton’s Root Mean Squared Propagation unpublished algorithm as the

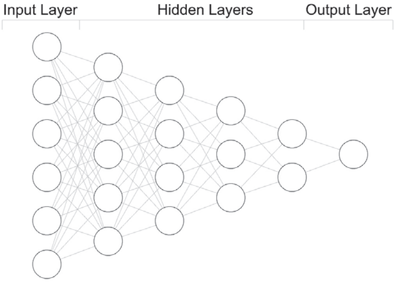
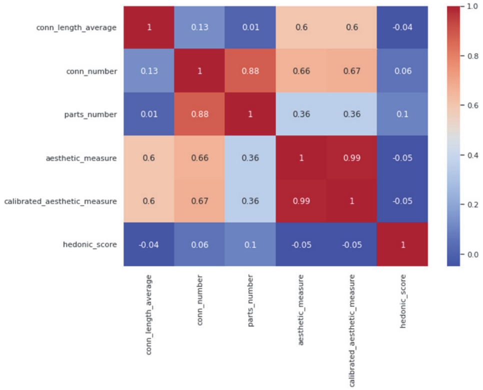


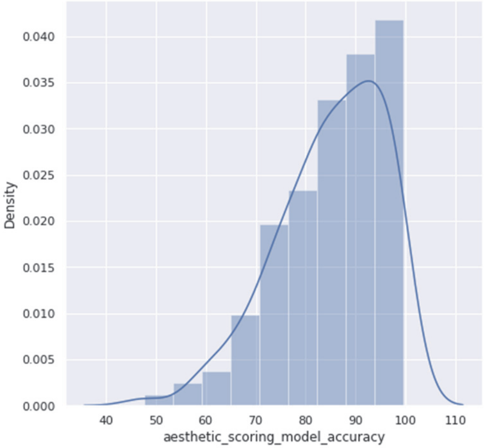
Figure 6
Diagram of the Artificial Neural Network



	conn_length_average	conn_number	parts_number	aesthetic_measure	calibrated_aesthetic_measure	hedonic_score
conn_length_average	1	0.13	0.01	0.6	0.6	-0.04
conn_number	0.13	1	0.88	0.66	0.67	0.06
parts_number	0.01	0.88	1	0.36	0.36	0.1
aesthetic_measure	0.6	0.66	0.36	1	0.99	-0.05
calibrated_aesthetic_measure	0.6	0.67	0.36	0.99	1	-0.05
hedonic_score	-0.04	0.06	0.1	-0.05	-0.05	1

1.0
0.8
0.6
0.4
0.2
0.0

Table 2
Correlation Matrix of all inputs and target output of the ANN



Graph 2
Histogram of the Accuracy of the ASM

optimiser and mean_squared_error as a loss. This model has an 85% accuracy rate.

RESULTS AND DISCUSSION

The paper is part of more extensive research for developing a computational framework for quantifying the architectural aesthetic experience. With the proliferation of computers in architecture, designers rely on CAD tools to generate and evaluate forms under various criteria like environmental performance, cost and profit optimisation, and structural behaviour. One of the goals of this research is to empower designers with tools to assist them to make design decisions also based on quantitative aesthetics.

Utilising social media platforms' feedback from specific audiences proved to be a powerful tool to train neural networks and calibrate aesthetic measure formulas. It can be employed to target particular audiences interested in specific architectural projects being developed currently or in the future. Therefore, the bias inherent in approaching a limited group of people is productive for defining an average aesthetic judgement for this specific group. That being the case, the problem of aesthetic evaluation as an evolutionary tract is irrelevant to the context of this paper.

Both CalAM and ASM proved to be extremely fast, outputting scores for new images in seconds. This responsiveness allows designers to test various options and be assisted by the crowdsourced trained or calibrated methods. It is possible to imagine such methods being applied to navigate solution spaces with evolutionary algorithms and implement them in optioneering systems to supplement economic, structural and environmental criteria with the final users' presumed aesthetic judgment.

All sub-methods should be integrated into a single CAD environment to be tested. At the current stage, each sub-method operates on different software, requiring its manipulation by a user on each step. MSER and parts-from-wholes run on our C#.net custom software entitled Aesthetics Framework. The aesthetic measure formula runs on Grasshopper. The ASM was trained and runs on Google Colab.

As defended above, our methods focus only on image composition. One could argue that architecture is a spatial problem. The authors respond that this could be compensated by having our methods applied to a series of images on movement, like CGI animations or video footage.

The power of composition as the basis of aesthetic judgement is that designers manipulate them as a good portion of their professional activity, especially when dealing with aesthetics. However, the larger audience's architecture assessment also involves other formal issues, like colour, light, texture, and proximity. Our brains constantly recognise objects like human figures, faces, and things. To leave alone other senses besides vision that are not part of the scope of our research.

If the presented accuracy is insufficient, we could incorporate more models to judge images, like Semantic Segmentation, Object Recognition and counting, colour and brightness histograms, and spatial semantics, all currently being done with CV. Recent developments in ML have been pointing to combining various models to increase accuracy (Jumper et al, 2021). In the next step, it is possible to implement continuous crowdsourcing of the hedonic response of the audience to further adjust the models.

A future experiment incorporating the workflow of (1) crowdsourcing to a specific audience the aesthetic judgement of images, (2) training a model for estimating the score of new images, (3) applying this model to navigate a solution space and evaluate multiple new design options, (4) crowdsourcing to the same audience the judgement of these design options and, finally, (5) compare the judgement from step 3 to step 4, is the necessary future work to prove the effectiveness of these methods in practice.

CONCLUSION

Both methods introduced proved to fit their prediction to the AHS with a certain accuracy. It is vital to notice that each architectural project has a specific audience and requires gathering new AHS to recalibrate CalAM and retrain ASM. In our test, the CalAM reached 66.98% accuracy and the ASM 85% accuracy. It is crucial to notice that the worst accuracy in our tests for CalAM is 8%, and it reaches a density peak at 75% accuracy (Graph 1). ASM has a minimum accuracy of 50%, and the density peak is at 95% accuracy (Graph 2).

Calibrating CalAM using Galapagos took 3 minutes on a quad-core Core i7-11800H @ 2.30GHz with 16GB RAM, and training 3000 epochs for the ASM took 2 hours using Google Colab server GPUs. Computing estimates for quantitative judgement of new images are in real-time for both methods.

Considering the accuracy improvement and that training is only necessary one time for each audience, it is highly recommended to apply the ASM compared to CalAM.

Crowdsourcing audiences of a project to train computational quantitative aesthetics models and formulas is a promising method for designing a more diverse and inclusive built environment.

REFERENCES

Bense, M. (1965). *Aesthetica. Einführung in die neue Ästhetik*, Baden-Baden: agis.

Birkhoff, G. D. (1933). *Aesthetic Measure*. Boston: Harvard University Press.

Dubey, A., Naik, N., Parikh, D., Raskar, R., Hidalgo, C. A. (2016) ´Deep Learning the City: Quantifying Urban Perception At A Global Scale´ in arXiv:1608.01769v2

Jumper, J., Evans, R., Pritzel, A. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

Kiemle, M. (1967). *Ästhetische Probleme der Architektur unter dem Aspekt der Informationsästhetik*, Schnelle, Qickborn.

Maser, S. (1968) Numerische Ästhetik.

Matas, J. (2004). ´Robust wide-baseline stereo from maximally stable extremal regions´ In Image and Vision Computing Volume 22, Issue 10, 1 September 2004, Pages 761-767

Salesses, P., Schechtner, K., Hidalgo, C. A. (2913) The Collaborative Image of The City: Mapping the Inequality of Urban Perception PLoS ONE 8(7): e68400. doi:10.1371/journal.pone.0068400

Sardenberg, V., Burger, T. and Becker, M. (2019). ´Aesthetic Quantification as Search Criteria in Architectural Design – Archinder´ in 37th eCAADe and 23rd SIGraDi Conference

Seresinhe, C. I., Precis, T. and Moat, H.S. (2017). ´Using deep learning to quantify the beauty of outdoor places´ in R. Soc. open sci. 4: 170170. Http://dx.doi.org/10.1098/rsos.170170

Shimamura, A. P. (2014). 'Towards a Science of Aesthetics: Issues and Ideas', in Shimamura, AP and Palmer, SE (eds) 2014, *Aesthetic Science*, Oxford University Press, pp. 3-28

Verma, D., Jana, A. and Ramamritham, K. (2018) ´Quantifying Urban Surroundings Using Deep Learning Techniques: A New Proposal´ in Urban Sci. 2018, 2(3), 78; https://doi.org/10.3390/urbansci2030078