



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE  
ESCUELA DE INGENIERÍA

**DEEP NEURAL NETWORK MODELS  
WITH EXPLAINABLE COMPONENTS FOR  
URBAN SPACE PERCEPTION.**

**ANDRÉS CÁDIZ VIDAL**

Thesis submitted to the Office of Research and Graduate Studies  
in partial fulfillment of the requirements for the degree of  
Master of Science in Engineering

Advisor:  
**HANS LÖBEL**

Santiago de Chile, July 2020

© MMXX, ANDRÉS CÁDIZ VIDAL



PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE  
ESCUELA DE INGENIERÍA

**DEEP NEURAL NETWORK MODELS  
WITH EXPLAINABLE COMPONENTS FOR  
URBAN SPACE PERCEPTION.**

**ANDRÉS CÁDIZ VIDAL**

Members of the Committee:  
**HANS LÖBEL**  
**PATRICIO DE LA CUADRA**  
**DENIS PARRA**  
**JORGE PÉREZ**

Thesis submitted to the Office of Research and Graduate Studies  
in partial fulfillment of the requirements for the degree of  
Master of Science in Engineering

Santiago de Chile, July 2020

© MMXX, ANDRÉS CÁDIZ VIDAL

*Gratefully to everyone that has  
supported me over the last 7 years*

## **ACKNOWLEDGEMENTS**

Thanks to my family and girlfriend for their unmeasurable support during the time this research was done. Thanks to my advisor for his great guidance and contributions. Thanks to all the members of *IALAB PUC* and *Instituto Milenio Fundamentos de los Datos* project #4 for all the very useful feedback and conversations. Thanks to ANID (former CONICYT) for granting me the National Masters scholarship, which enabled me to complete my studies without any issues.

## TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF FIGURES	viii
LIST OF TABLES	xi
ABSTRACT	xii
RESUMEN	xiii
1. INTRODUCTION	1
2. RELATED WORK	4
2.1. Understanding and quantifying urban perception.	4
2.1.1. Classic approaches.	4
2.1.2. Pure machine learning approaches.	5
2.1.3. Mixed approaches.	7
2.2. Explainability in machine learning.	9
3. DATASET	13
3.1. Description	13
3.2. Analysis and preprocessing	13
4. PROPOSED MODEL	16
4.1. Problem Definition	16
4.2. Network architectures	17
4.2.1. SegRank base.	17
4.2.2. SelfSegRank	20
4.2.3. AttentionSegRank	22
4.3. Additional components	23
4.3.1. Loss function	23

4.3.2. Semantic Dropout . . . . .	25
4.4. Baselines . . . . .	26
4.4.1. ResNet50 + MLP . . . . .	26
4.4.2. ResNet50 + Self Attention layer + MLP . . . . .	26
5. METHODOLOGY . . . . .	28
5.1. Models and training . . . . .	28
5.2. Visualization . . . . .	29
6. RESULTS . . . . .	30
6.1. Quantitative results . . . . .	30
6.1.1. Model performance . . . . .	30
6.1.2. Training behavior . . . . .	31
6.2. Visualization results . . . . .	33
7. DISCUSSION . . . . .	38
7.1. Effect of semantic segmentation on learning . . . . .	38
7.2. Segmentation as an attention subject. . . . .	38
7.3. Relationship between urban perception and semantic segmentation . . . . .	41
7.4. Effect of combining features and segmentation. . . . .	45
7.5. Effect of attention over semantic segmentation on model explainability. . . . .	48
8. CASE STUDY . . . . .	51
9. CONCLUSIONS . . . . .	55
9.1. Contribution to the state of the art. . . . .	55
9.1.1. Enhancing explainability through high level features and attention. . . . .	55
9.1.2. Semantic segmentation as part of a neural network. . . . .	56
9.2. Future research directions. . . . .	56
9.2.1. Improving dataset quality. . . . .	56
9.2.2. Additional high level features. . . . .	57
9.2.3. Improved attention mechanisms. . . . .	57

REFERENCES	58
APPENDIX	66
A. Semantic Segmentation . . . . .	67
A.1. Visual representation . . . . .	67
A.2. Segmentation distribution . . . . .	68
B. Significance Tables . . . . .	69
C. Correlation Tables . . . . .	71

## LIST OF FIGURES

2.1	Place pulse 2.0 survey . . . . .	6
2.2	Beta Coefficients . . . . .	8
2.3	Attention on VQA . . . . .	12
3.1	Repetition histogram . . . . .	14
4.1	Example of Semantic Segmentation . . . . .	18
4.2	PSPNet architecture . . . . .	19
4.3	First model architecture . . . . .	20
4.4	Attention Mechanism . . . . .	21
4.5	Self Attention network . . . . .	22
4.6	Self Attention Model output . . . . .	22
4.7	Segmentation as key network . . . . .	24
4.8	Effect of Semantic Dropout . . . . .	25
4.9	ResNet + SelfAttention . . . . .	27
6.1	ResNet Training curves . . . . .	31
6.2	ResNetAttn Training curves . . . . .	32
6.3	SegRank Training curves . . . . .	32
6.4	SelfSegRank Training curves . . . . .	33
6.5	AttentionSegRank Training curves . . . . .	33
6.6	Multi attribute Visualization Sample . . . . .	34

6.7	Attention by segmentation class heatmaps.	34
6.8	Attention by segmentation class heatmaps.	35
6.9	Result samples	36
6.10	Result samples	37
7.1	Wealthy Training curves	39
7.2	Percentage of class significance when total	40
7.3	Percentage of class significance when present	41
7.4	Example of significance ratio	42
7.5	Correlation between segmentation and attention percentages for each class and quantified scores.	43
7.6	Correlation comparison for significant images	44
7.7	Score vs segmentation scatter plots	44
7.8	Correlation comparison	46
7.9	SelfSegRank Correlation	47
7.10	SelfSegRank score vs segmentation scatter plots	47
7.11	Vegetation attention averages	49
7.12	Road attention averages	50
8.1	Urban Perception for Santiago de Chile	52
8.2	Wealthy score vs wealth distribution of Santiago de Chile	53
8.3	Poverty rate vs perceived wealthy score	53
8.4	Vulnerability index vs perceived depressiveness score	54

A.1	Segmentation color palette	67
A.2	CityScapes sample	67

## LIST OF TABLES

3.1	Votes Distribution . . . . .	15
5.1	Hyper parameters . . . . .	28
6.1	Model accuracy . . . . .	31
A.1	Segmentation Distribution . . . . .	69
B.1	Percentage of class significance . . . . .	70
B.2	Percentage of class significance when present . . . . .	71
C.1	Score and segmentation correlation. . . . .	72
C.2	Score and attention correlation. . . . .	73

## ABSTRACT

Urban perception has been an important research subject for at least 60 years, with studies being conducted by many different disciplines, using a variety of methodologies mainly based on surveys over either real or simulated urban environments. The recently increased availability of large amounts of data and highly scalable data collection methods powered by the modern web has allowed for new techniques from other domains to be extended to the estimation of urban perception. In particular, machine learning methods used as either stand alone models or feature extraction tools have proven to be very effective for automatic quantification of the perception. This methods (neural networks in particular) present the disadvantage of having a black box nature, which can make it hard to understand the obtained results from a human point of view, therefore limiting their application.

In this work we present a novel neural network architecture for automatic urban perception quantification. Based on an image, our best model, named AttnSegRank, can output an estimated urban perception score, along with a set of weights (displayable as a heatmap) that reflect the importance of each part of the image on the calculation of the score. It achieves this by including the output of a pretrained semantic segmentator leveraged with an attention mechanism as part of the architecture. The model we show in this work presents very similar performance with those in the previous literature but with a much better interpretability, making it not only a more useful model for urban perception measuring and research, but a contribution to explainability in the deep learning and computer vision fields that can be applied to other tasks as well.

**Keywords:** urban perception, deep learning, explainable artificial intelligence.

## RESUMEN

La percepción urbana ha sido un tema de investigación importante por al menos 60 años, con trabajos hechos desde una variedad de disciplinas, usando distintas metodologías principalmente basadas en encuestas sobre ambientes urbanos reales o simulados. Recientemente, el surgimiento de las tecnologías web ha aumentado masivamente la disponibilidad de grandes volúmenes de datos y de técnicas de recolección de datos altamente escalables permitiendo que técnicas pertenecientes a otros dominios sean utilizadas para estimar la percepción urbana. En particular, algunos métodos de aprendizaje de máquina, usados ya sea como modelos completos o como herramientas de extracción de características, han demostrado ser muy efectivos para la cuantificación automática de la percepción. Estos métodos (en particular las redes neuronales) presentan la desventaja de tener una naturaleza caja negra, que dificulta la capacidad de entender los resultados obtenidos desde el punto de vista humano, lo que limita su aplicabilidad.

En este trabajo presentamos una nueva arquitectura de red neuronal para la cuantificación de la percepción urbana. A partir de una imagen nuestro mejor modelo, *AttnSeg-Rank*, entrega como resultado una estimación de la percepción, junto a un conjunto de pesos que (visualizables como un mapa de calor) que reflejan la importancia de cada parte de la imagen en el cálculo del score. Esto se lo logra incluyendo un segmentador semántico combinado con un mecanismo de atención como parte de la arquitectura de la red. El modelo que mostramos en este trabajo logra un rendimiento similar a los existentes en la literatura pero con mucho mejor interpretabilidad, haciéndolo no solo un modelo más útil para medición e investigación de la percepción urbana, si no que una contribución a la explicabilidad en aprendizaje profundo y visión por computador que se puede aplicar a otras tareas.

**Palabras Claves:** percepción urbana, aprendizaje profundo, inteligencia artificial explícable.

## 1. INTRODUCTION

*Urban perception* is a feeling held by people about a location. These feelings can be and are often related to a particular characteristic, like happiness or beauty, or also inherently negative ones, like insecurity or fear (Ordonez & Berg, 2014). Understanding the cause of these feelings is a complex task, since unique social and psychological aspects of each individual affect how they perceive and the spaces they observe (Nasar, 1990).

Visual urban perception is responsible for a large part of the experience that people go through while being at or using an urban space. This not only affects how much the spaces themselves are used (Khisty, 1994) but also the use of related means of transport (Antonakos, 1995). Other studies have also found correlations between urban perception, crime statistics (Ordonez & Berg, 2014) and wealth, and therefore used it as a proxy measure of inequality (Ordonez & Berg, 2014; Salesses, Schechtner, & Hidalgo, 2013; Rossetti, Lobel, Rocco, & Hurtubia, 2019).

Additionally, being able to measure a community's need and perception of a city at scale is something of key importance on developing cities. The insights can be applied for design of public policy so that local governments can allocate their resources more efficiently (Santani, Ruiz-Correa, & Gatica-Perez, 2018).

Traditional methods for obtaining this information, consist of hand made polls about specific locations, making systematic quantification of perception an extremely costly and hard to escalate task (Nasar, 1990; Clifton & Ewing, 2008). An alternate approach consists of surveys based on computer generated images of simulated spaces. This scheme is more scalable, but is limited to experimental design and cannot be directly applied to real urban spaces (Laing et al., 2009; Iglesias, Greene, & Ortúzar, 2013).

Currently, thanks to the great volumes of data generated by web platforms (Salesses et al., 2013) and to modern deep learning (DL) and computer vision techniques (LeCun, Bengio, & Hinton, 2015), new solutions for quantifying urban perception at scale have

become feasible. The Place Pulse 2.0 dataset (Dubey, Naik, Parikh, Raskar, & Hidalgo, 2016), is the most significant example of this, consisting of pairs of images along with labels that indicate which of the images is more representative of a particular attribute. Previous studies have achieved significant results with it, either by applying traditional deep learning (Dubey et al., 2016) or by combining it with other approaches (Rossetti et al., 2019; F. Zhang et al., 2018). In general, works trying to quantify urban perception at scale consist of training deep convolutional neural network models (DCNN) (LeCun et al., 1989) with datasets of urban images that have some sort of label that is used as an estimator for the perception of that urban space, such as Place Pulse 2.0.

However, current deep learning methodologies, have the disadvantage of being “black boxes”. In other words, they lack a direct or systematic way to explain or interpret the obtained results. This problem comes from the layered structure of the neural network models and from the millions of learnable parameters they contain. Many of the problems in which these models are used would greatly benefit of more human understandable explanations of the results, since they provide more confidence and control over the decisions influenced by the systems. Due to that, explainability has become a very important area of research for the deep learning field (Adadi & Berrada, 2018; Ras, van Gerven, & Haselager, 2018). For the particular case of urban perception, explainability of the results is of utmost importance, since the added information is valuable for the design of public policy. For instance, it could be used to better discriminate which locations would be better recipients of an intervention, and which elements to modify so it convenes an effective improvement of perception. Despite that, a fully explainable approach is yet to be proposed and the purpose of this work is to progress towards that objective.

The research community has realized the importance of explainability and has taken the research in two main directions. One direction is to design novel neural network architectures and training methods with the intention of making them interpretable, such

as the work by Dong, Su, Zhu, and Zhang (2017). The other direction is to create post-hoc algorithms (Adadi & Berrada, 2018) that analyze the results given by the neural network. These algorithms commonly use machine learning models, including neural networks (Ghorbani, Wexler, Zou, & Kim, 2019). In particular, the work by Rossetti et al. (2019), presents an approach that uses semantic segmentations of images (Badrinarayanan, Kendall, & Cipolla, 2015) as input for a discrete choice model that estimates an utility function quantifying the perception of citizens regarding different concepts, such as beauty and safety. Among other things, this approach allows for a post-hoc aggregated analysis of the results, based on the coefficients of the utility functions, which quantify the importance of each of the explainable input variables. It is important to note that this type of techniques usually imply a trade-off between the prediction performance of the model and it's explainability.

The objective of this work is to design and train a model for the urban perception problem, that can give explainable insights on an instance level. For that it proposes a novel solution, consisting of a neural network architecture, that is end-to-end trainable and by using semantic segmentation (Zhao, Shi, Qi, Wang, & Jia, 2016) and self attention mechanisms (Vaswani et al., 2017) can show explainable insights for each of the input images.

The remainder of this manuscript is organized as follows, chapter 2 summarizes relevant previous research. Chapter 3 describes the dataset used and how it was pre processed. In chapter 4 the problem is formally defined and our models are described. Chapter 5 gives details on model implementation and training. Chapter 6 presents the research results and 7 the related discussion. Finally, Chapter 8 consists of an applied case study on the city of Santiago Chile and chapter 9 gives the final conclusions and future research directions.

## **2. RELATED WORK**

This chapter consists of two sections. The first one shows an overview some of the different methods that have been previously used in the literature for understanding or quantifying urban perception. The second section summarizes the main aspects of the research on explainability on deep learning, and describes some techniques that have been applied in urban perception or other domains that are relevant for this work.

### **2.1. Understanding and quantifying urban perception.**

#### **2.1.1. Classic approaches.**

Methods for measuring perception of urban spaces have been part of the literature of several disciplines for many years, with some of the most influential studies dating back to 1960 (Lynch, 1960). Due to technological limits the literature consisted mainly of different types of qualitative surveys for a long time. These surveys consisted in having subjects complete different tasks such as drawing maps of a certain place (Lynch, 1960), evaluating fundamental aspects of a neighborhood (Nasar, 1990) or in more recent approaches, evaluating the impact of transformations generated with edited images (B. Jiang, Mak, Larsen, & Zhong, 2017). Most of these surveys were conducted in person or by phone, and then the results were analyzed manually, making it very difficult and costly to scale to multiple locations or larger amounts of samples. The main benefit of this approach, is that it allows for a precise control of the observation process since both the subjects being interviewed and the spaces in question are chosen by the researcher. Furthermore, the experiments conducted in person allow for the observer to use senses different than vision to analyze the subject space, resulting in a richer appreciation.

A different methodology, more common in economics and engineering, consists of using discrete choice models and stated choice surveys to model the effect of different variables in perception or other urban related variables (Rose & Bliemer, 2009; Iglesias et al., 2013; Torres, Greene, & Ortúzar, 2013). The amount and complexity of the variables

measured depends on the model design. To have and exact control of the variables that have an effect on the survey, computer generated images of urban spaces can be used (Iglesias et al., 2013; Torres et al., 2013).

The advantage of this method is that through the estimated parameters of the model, the effect of each of the studied variables on the perception estimation can be measured, allowing for quantitative results and an understanding of the impact different elements have on the perception of the urban landscape. The main disadvantage of this approach comes from the difficulty of the survey design; variables need to be chosen carefully and the process is vulnerable to biases from the model designer.

### **2.1.2. Pure machine learning approaches.**

Thanks to the massive adoption of web and mobile technologies such as Google Maps, new types of data are available in considerably large volumes, and new highly scalable ways of generating data can be designed and implemented quickly. These facts allows the application of data-intensive machine learning algorithms to new problems, including urban perception estimation. Several different datasets have been proposed for this problem, most of them based on surveys over large amounts of urban images (Salesses et al., 2013; Dubey et al., 2016; Quercia, O'Hare, & Cramer, 2014; Liu, Silva, Wu, & Wang, 2017; Santani et al., 2018). The most used, all consisting of pairwise comparisons of street view images, are *Place pulse 1.0* (PP 1) (Salesses et al., 2013) with measures of safety, class and uniqueness over images of 4 cities, *Urban Gems* with measures of beauty, quietness and happiness over images of London and *Place pulse 2.0* (PP 2) (Dubey et al., 2016), the largest dataset available, with measures of six different attributes over images of 56 different cities. All of these were collected through public online surveys of large scale, where the users are asked to choose the image most representative of an attribute of a pair, see Figure 2.1 for an example.

Earlier attempts at using these data for training models focused on turning the perception quantification into a classification problem by first ranking the images from the votes

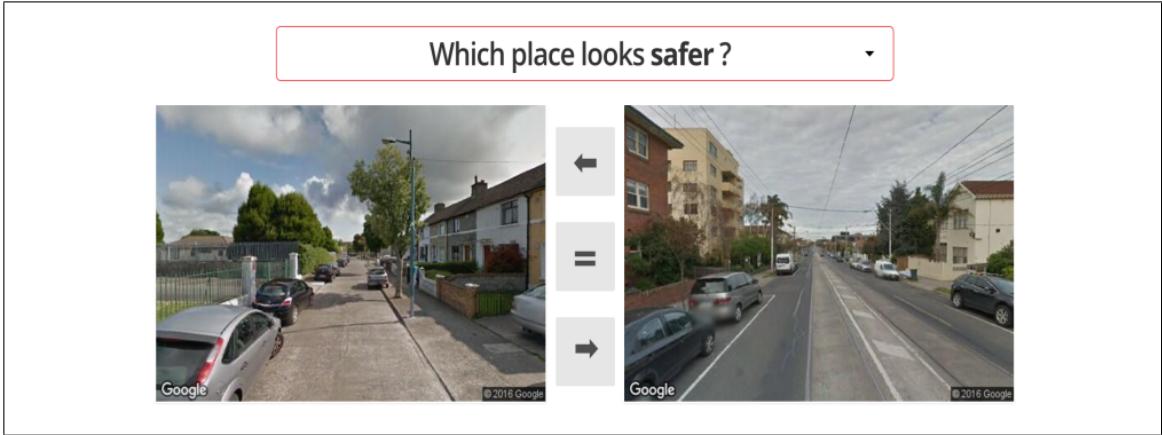


Figure 2.1. Snapshot of the place pulse 2.0 survey. Extracted from Dubey et al. (2016)

with manually engineered methods, such as the one suggested in Salesses et al. (2013), and then using the rank to split the data in two halves with a different label. Ordonez and Berg (2014) use this approach to train SVM models on PP 1 using different types of visual features as input, including features generated by a deep neural network (Donahue et al., 2014). On the PP 2 paper, the authors present the first end to end deep learning model for urban perception regression, which uses a typical transfer learning technique (Pan & Yang, 2010), consisting of an Imagenet (Deng et al., 2009) pretrained network for the base of the model, which is used as input for by two parallel modules, one for classification and one for regression. They train the architecture separately on the 6 different attributes of the dataset and the models learn to emulate human voting and to output a urban perception score (through the regression module) on the image for the correspondent attribute. Other works (Porzi, Rota Bulò, Lepri, & Ricci, 2015; Santani et al., 2018) take similar approaches but pretrain models or use features based on the places dataset (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014), which provides better performance according to their results.

F. Zhang et al. (2018), train models on PP 2 by combining a DCNN features and an SVM classifier to obtain perception indicators of Beijing. They also use a semantic segmentation model (Cordts et al., 2016) on the images and use the results as input to a

linear regression, interpreting the regression weights as an indication of importance of the different segmentation classes on perception. On a following work (F. Zhang et al., 2020) they train one deep network to predict all 6 attributes of PP 2 in one forward pass, they do this using an end-to-end architecture similar to Dubey et al. (2016) but adding one output and loss component for each attribute.

It is important to note that most of the literature so far is more focused on applying the models to new cities (F. Zhang et al., 2018; Santani et al., 2018; Costa, Soares, & Marques, 2019; Rossetti et al., 2019) or generating new datasets with new attributes (Santani et al., 2018; F. Zhang et al., 2020), than it is on improving model design and performance. This is consistent with the fact that so far no good measures of performance for this problem have been defined, due to the fact that the datasets do not provide a measure of perception per se but a proxy through the survey votes. The objective of the models in the literature is to rank the images by the estimated perception of an attribute, but they measure performance using accuracy on classification of the human votes, which does not necessarily correlate with the models capacity to generalize and rank well, especially in conflicted cases where even human voters would have difficulties (F. Zhang et al., 2018). Despite the fact that models in the literature do not surpass 70% classification accuracy on PP2, the actual ranking task seems to have correct results either by visual inspection, or by comparing with metrics from other domains such as crime rates or wealth indicators (Rossetti et al., 2019; F. Zhang et al., 2018; Ordonez & Berg, 2014).

### **2.1.3. Mixed approaches.**

With the intention of generating more or different insights, usually through more explainability, some work in the literature consists of combinations of computer vision or machine learning methods with other techniques. Rossetti et al. (2019) use a combination of low-level high-level features of the images as input for a discrete choice model that estimates perception. They extract low-level features with traditional computer vision methods like edges or blobs and the high-level features with a pretrained neural network

for semantic segmentation. As the semantic segmentation features allow for a posthoc analysis of the results, the authors are able to reach conclusions like "Images with more sidewalks were deemed to be safer, livelier and wealthier, but less beautiful on average" and they present a table with the significance of each of the segmentation classes in each of the six PP 2 attributes according to the discrete model parameters. On a similar line, as was mentioned earlier, F. Zhang et al. (2018) in addition to their main method use semantic segmentation features (they aggregate them by percentage of pixels on the image) as an input for multivariate linear regression allowing for similar conclusions to those of Rossetti et al. (2019) but using the beta coefficients (see Figure 2.2).

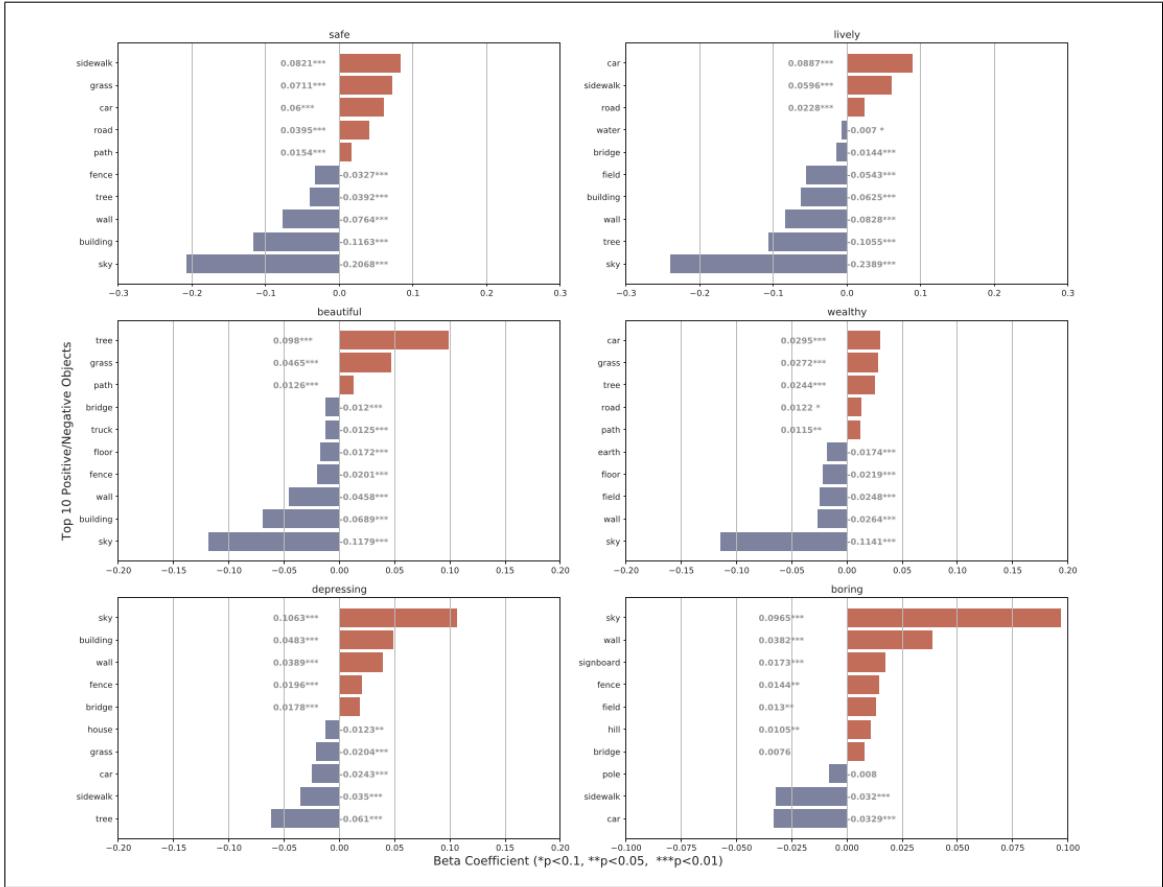


Figure 2.2. Linear regression beta coefficients for most significant objects.  
Extracted from F. Zhang et al. (2018)

On another work Seresinhe, Preis, and Moat (2017) train a DCNN to calculate the beauty of outdoor images, using transfer learning from the Places dataset, but separately they use a model trained on that same dataset to obtain text tags from the scenes such as 'Mountain' or 'Tower', and similarly to F. Zhang et al. (2018) they use a regression model (elastic net) to make conclusions about the significance of the concepts on the perception of beauty. The disadvantage of these approaches is that they provide more insights of the results only at a general level, and therefore do not allow for conclusions on a per instance level, which is what this work intends to do.

Authors of Costa et al. (2019) perform an agreement analysis for this type of dataset, by building their own dataset of pairwise comparisons for the perception of safety, but used it for generating clusters of users based on the semantic segmentation of the images they voted for. They conclude that the existence of most clusters is due to a lack of enough comparisons to do a good characterization and that given enough votes all users converge to one generic profile. It is important to note that authors don't provide any social or demographic information of the 439 users that participated in the survey, and no other similar studies have been done so far so their conclusion hasn't been replicated.

## **2.2. Explainability in machine learning.**

As was mentioned before, explainability has become a very active area of research in machine learning, due to the large increase in the usage of ML models for different day to day applications that affect the life's of thousands of people (Ras et al., 2018). For example, in cases where model outputs are used for analytics or decision making, explainability can make the model both more trustworthy and informative.

Adadi and Berrada (2018) summarize the reasons for enhancing explainability in four points:

- (i) Explain to justify: To fulfil the need for reasons of a particular ML generated outcome.

- (ii) Explain to control: To allow a better handling of model behavior.
- (iii) Explain to improve: The additional understanding of model outputs is useful to design improvements on the systems.
- (iv) Explain to discover: As a model overcomes human performance in a task, if its doing so in an explainable manner, then new knowledge for humans may be obtainable.

It is also important to note that laws and regulations related to this topic may become norm in the future such as with the *European Union General Data Protection Regulation (GDPR)* (2016). According to it's articles 13, 14 and 15, when personal data is collected for automated decision-making, the subject has the right to access, and the data controller is obliged to provide, “meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject”, which will be very difficult to comply with, when working with something like a black box neural network.

One of the two most common approaches to explainability in the literature are post-hoc methods (Adadi & Berrada, 2018) which try to obtain insights about how the models work, after the process of inference over all the dataset is completed. The methods mentioned in section 2.1.3 are examples of this approach. Other more complex methods found in the recent literature are based on analyzing model sensitivity to semantically meaningful concepts on input (Kim et al., 2017; Shi, Zhang, Wang, & Reddy, 2020), concepts that may be automatically mined from the data as in the approach proposed by Ghorbani et al. (2019). These techniques, although very promising, are still too recent and are not extensible to many domains.

The other approach, which is the one followed by this work, consists of taking advantage of the model design to improve interpretability. This can be done by either using existent features of the model or by introducing architectural changes that make them more explainable. A traditional example of this approach are rule based models like decision

trees (Breiman, Friedman, Stone, & Olshen, 1984). Due to the black box nature of deep neural networks this becomes a much more complex task for deep learning, and its an important area of research.

Earlier solutions found in the literature consist of augmenting model input with semantic information, such as text, object bounding boxes or even knowledge bases (Dong et al., 2017; Zhuo, Cheng, Zhang, Wong, & Kankanhalli, 2019; G. Li, Wang, & Zhu, 2019). These methods usually require additional supervision which restricts them to densely annotated datasets such as Visual Genome (Krishna et al., 2016), and the way the neural networks actually use the additional information is not always clear.

Another technique in the literature is the use of attention-based models (Bahdanau, Cho, & Bengio, 2014), which have layers that consist on using a part of the input (usually called query) to compute a set of weights for the rest of the input (usually called value). In other words the query matrix is used to determine which vectors of the value matrix are more relevant for the calculation of the output. The attention weights are usually computed through a linear transformation and a softmax operation on the query, giving them the property of being a probability distribution over the vectors of the value matrix (Cordonnier, Loukas, & Jaggi, 2019), which is used to increase or decrease, parts of the value matrix and therefore the layer output. A particular case of attention is self-attention, which means that the same vector is used as both query and input. A common attention architecture in the recent literature is the transformer (Vaswani et al., 2017), which has been widely adopted in both language and vision tasks (Devlin, Chang, Lee, & Toutanova, 2018; Radford et al., 2019; Bello, Zoph, Vaswani, Shlens, & Le, 2019; L. H. Li, Yatskar, Yin, Hsieh, & Chang, 2019; Carion et al., 2020).

Attention models not only have provided improvements in performance, but also have the additional value that the weights can be used to interpret what the network is doing, providing explainable information about the model’s decision process for each data instance (Wiegreffe & Pinter, 2019). Clark, Khandelwal, Levy, and Manning (2019) analyse the attention outputs of the NLP transformer model BERT and show how they correspond

well to linguistic notions of syntax, Y. Zhang, Niebles, and Soto (2019) create an explainable VQA model by adding supervised self-attention layers and visualizing their output as heatmaps over the input images (see Figure 2.3). Cordonnier et al. (2019) present a theoretical relationship between self attention and convolutional layers, and as part of their work they provide an interactive visualization of the attention weights<sup>1</sup>. M. Jiang, Chen, Yang, and Zhao (2020) present a dataset that includes attention labels for images generated by human eye movement, allowing models to learn correct human like attention patterns. These works show that visualization of attention weights has become a prominent tool for improving model explainability on the recent deep learning literature not only on language but on vision tasks as well (Z. Zhang, Lan, Zeng, Jin, & Chen, 2020; Johnston & Carneiro, 2020; Carion et al., 2020).

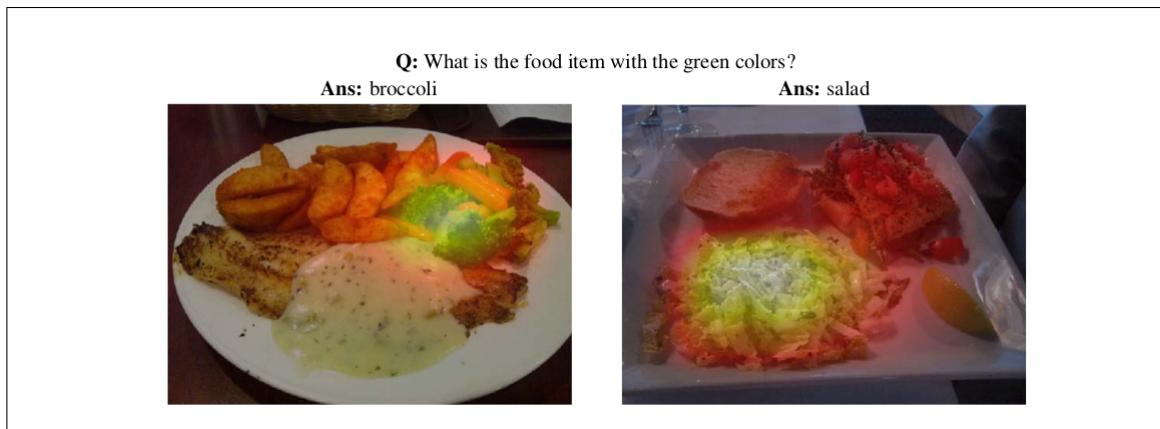


Figure 2.3. Visualization of attention weights for the VQA problem. Extracted from Y. Zhang et al. (2019)

---

<sup>1</sup>Available at [epfml.github.io/attention-cnn/](https://epfml.github.io/attention-cnn/)

### **3. DATASET**

This chapter presents an analysis of the dataset used. On Section 3.1 an overall description and main statistics are shown. Section 3.2 presents an early analysis of the data along with the preprocessing steps taken. Finally section 3 makes an analytical definition of the problem of learning urban perception from this data.

#### **3.1. Description**

As was mentioned previously, this work is based on the Place Pulse 2.0 dataset (Dubey et al., 2016). PP 2.0 is a crowdsourced dataset designed for learning urban perception from street view like images. Unlike regular datasets for supervised machine learning, that have labels for each image, Place Pulse consists of pairwise comparisons between images, and the ground truth is a vote representing which of the images is more representative of an attribute (ties are also possible). That structure makes traditional classification / regression approaches inapplicable, but opens the door for pairwise based ranking techniques, that are more suitable to urban perception since a ground truth for how much an image represents an abstract attribute such as "safety" it's impossible to define.

The dataset consists of approximately 1.2 million pairwise comparisons of 112,000 images from 56 cities, distributed on 6 attributes: wealthy, safety, depressing, boring, lively and beautiful, making it the biggest available dataset for urban perception. The crowdsourcing survey was active for 5 years and it was answered by 81,630 different users. Demographic information about the users was not collected.

#### **3.2. Analysis and preprocessing**

As a first preprocessing step all noisy images are removed by using a file size threshold, since files small in size are mostly google api errors or unintelligible places like dark tunnels.

It is important to note that, unlike most crowdsourced datasets, the authors of PP did not perform a validation on the votes. 99.59% of the image pairs that appear in the data set have a single vote in a category (see 3.1 for details), making it impossible to corroborate if they are reasonable by comparing the votes of multiple people. Even though previous research indicates that answers to this surveys aren't affected by user bias or demographics (Salesses et al., 2013; Costa et al., 2019), the inconsistency in the votes is a clear dataset disadvantage: 34% of the pairs that have more than one vote in an attribute show inconsistencies between the votes.

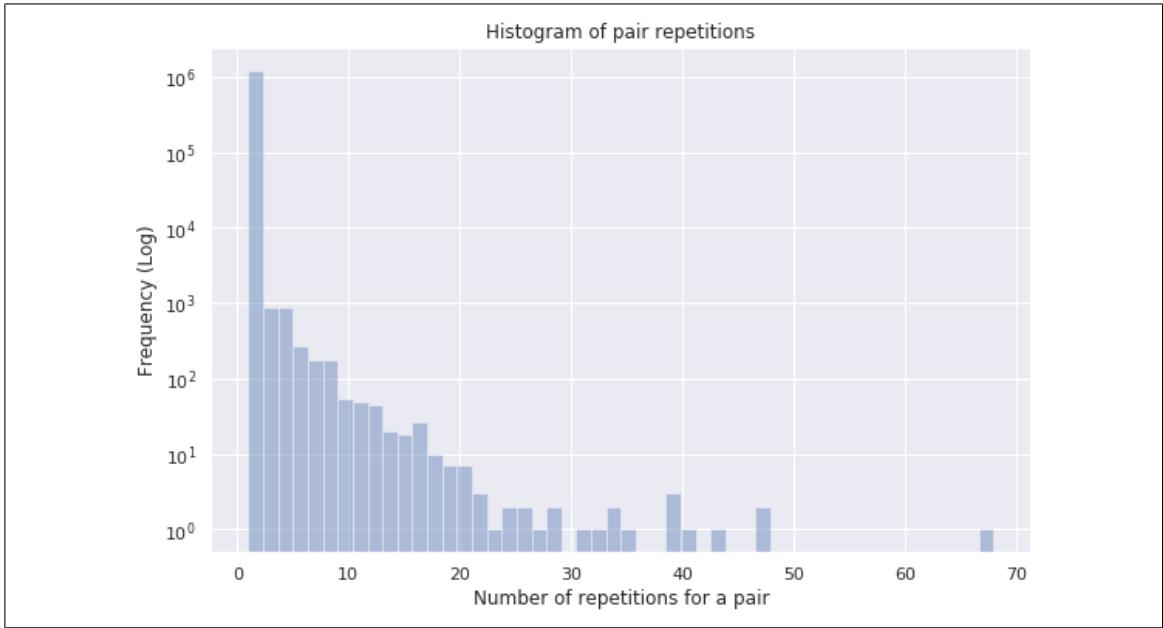


Figure 3.1. Histogram for amount of repetitions for each pair of images

For this work we completely remove all inconsistent duplicates and keep a single instance of those consistent. After these, steps 1,207,938 votes for 111,299 images are left. See Table 3.1 for the exact vote distribution.

Table 3.1. Vote distribution after preprocessing.

<b>Attribute</b>	<b># of votes</b>
Wealthy	150,370
Safety	364,130
Depressing	130,781
Boring	125,744
Lively	263,123
Beautiful	173,790
Total	1,207,938

Users of the survey had the possibility of voting that a pair is tied for an attribute, meaning that they didn't perceive any significant difference. Previous works usually discard this data and don't use it for learning, focusing only on the votes where a preference was chosen (Dubey et al., 2016; F. Zhang et al., 2018; Ordonez & Berg, 2014). After preprocessing 15.3% of the votes are ties, which means a significant amount of information is lost by disregarding them. Due to that we decided to add additional rules to the learning problem in order to be able to use these votes for learning. Details are shown on the following chapter.

## 4. PROPOSED MODEL

This chapter presents a detailed explanation of the neural network models proposed in this work and the correspondent baselines used for comparison. In Section 4.1 we give a formal definition of the learning problem. In Section 4.2 the architectures of the main networks are shown. In Section 4.3 we detailed additional components used on the models and training. Finally, Section 4.4 shows the baselines models used for the ablation study of both performance and explainability.

### 4.1. Problem Definition

Following a formulation similar to Dubey et al. (2016), each attribute  $A$  in the PP 2.0 consists of a set  $I_A$  of images and a set  $P_A$  of votes for those images. Each image in  $I_A$  is a tensor  $x \in \mathbb{R}^{h \times w \times 3}$  where  $h$  and  $w$  are the height and width respectively. Each vote in  $P_A$  is a triple  $(x_1, x_2, y)$  where  $x_1$  and  $x_2$  are images in  $I_A$  and  $y \in \{1, 0, -1\}$ . A triple  $(x_1, x_2, y) \in P_A$  represents a comparison where  $y = 1$  states a preference of  $x_1$  over  $x_2$ , and  $y = -1$  represents a preference of  $x_2$  over  $x_1$ . The value  $y = 0$  represents a tie.

The objective is to, for each attribute, learn a ranking function  $f_A : \mathbb{R}^{h \times w \times 3} \rightarrow \mathbb{R}$  that maps the image tensor to an urban perception score, satisfying the order given by the votes. Formally the maximum amount of the following constraints need to be satisfied:

$$y \cdot (f_A(x_1) - f_A(x_2)) > 0 \quad \forall (x_1, x_2, y) \in P_A \text{ and } y \in \{-1, 1\} \quad (4.1)$$

Unlike the previous literature, tie votes are also used in this work, generating the following additional constraints:

$$|f_A(x_1) - f_A(x_2)| < m \quad \forall (x_1, x_2, y) \in P_A \text{ and } y = 0 ; m \in \mathbb{R}^+ \quad (4.2)$$

Where  $m$  is a constant margin.

Since  $f_A$  is intended to learn a ranking of the input images, it is desirable that the function defines an order on the image space so that the ranking results are consistent. This condition can and should be enforced by model design (Köppel et al., 2019), but since the data is crowdsourced without validation, the constraints generated by Equation 4.1, do not represent a 100% transitive order. Because of that, it is infeasible for a model designed for ranking and therefore transitive by construction, to satisfy all of them. This issue makes it harder to obtain high scores in accuracy based metrics in practice, and those are the only ones available in the literature so far.

## 4.2. Network architectures

As was mentioned before, the main principle followed for model design is to enhance explainability while maintaining performance of the model as much as possible. With that in mind, we combine two state-of-the-art techniques from the deep learning literature, semantic segmentation and attention mechanisms, to design three novel architectures that present a significant improvement in explainability over traditional blackbox CNNs. We describe these architectures in the following sub sections, ordered by model complexity. It is important to note that for learning to rank on the PlacePulse dataset, two forward passes of the ranking network are required for each data instance (one for each image) and both scores are used for calculation of the loss. See section 4.3.1 for details.

### 4.2.1. SegRank base.

The traditional deep learning approach in computer vision, consists of using a pre-trained CNN (LeCun et al., 1989), on the Imagenet dataset (Deng et al., 2009), such as the ResNet (He, Zhang, Ren, & Sun, 2015), usually called the feature extractor, and then stacking a custom set of layers over its output features. Leaving the CNN weights fixed or updating them on training depends on the particular problem. This is the approach taken by most of the previous literature on urban perception (Dubey et al., 2016; Ordonez & Berg, 2014; F. Zhang et al., 2018).

In this work, we propose replacing the traditional feature extractors for a fully trained semantic segmentation network. The semantic segmentation task consists of assigning a label to every pixel in an image, and therefore it implies a fine grained detection of object edges, providing a rich amount of information that is human understandable. The output of a semantic segmentation model is a probability distribution over the different classes for each pixel, making it usable as a feature map of the image. See Figure 4.2 for an example.

We base our models on the PSPNet architecture (Zhao et al., 2016), since it is one of the highest performing models available in the literature. Its design is based on a ResNet50 and a pyramid pooling module, which consists on parallel poolings and convolutions at different scales, that are then concatenated and used to generate the output with a final convolution.

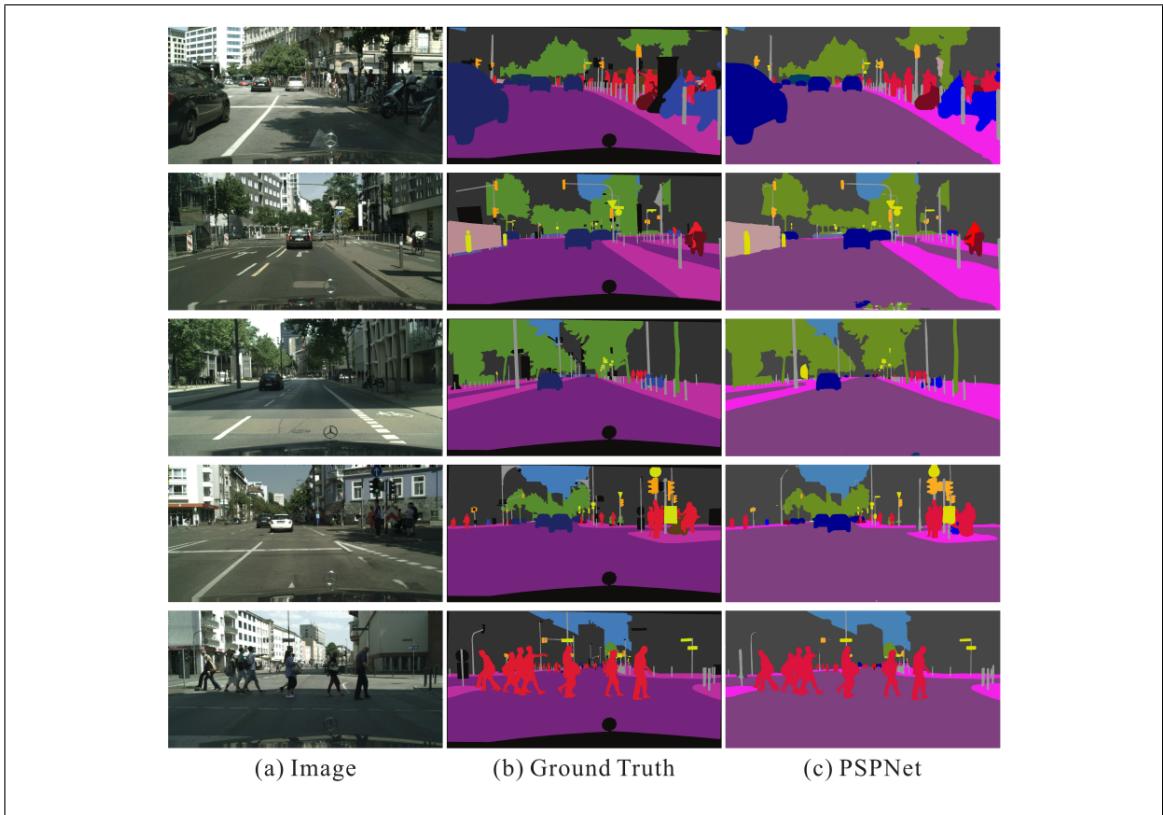


Figure 4.1. Examples of semantic segmentation by the PSPNet model on the CityScapes dataset. Extracted from Zhao et al. (2016)

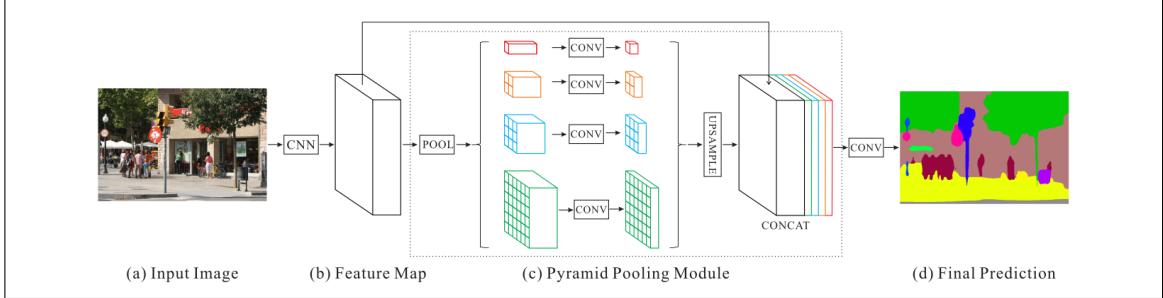


Figure 4.2. PSPNet architecture. Extracted from Zhao et al. (2016)

We train PSPNet on the CityScapes dataset (Cordts et al., 2016), since its urban images taken from a car have considerable similarity to Street View images, and its classes have proven informative for the urban perception problem in previous research (Rossetti et al., 2019; F. Zhang et al., 2018). After this process we keep the network weights fixed and use the output as features for subsequent layers. The segmentation output is a tensor  $S$ ,  $S \in \mathbb{R}^{h \times w \times C}$ , with  $C$  being the number of different classes. We experiment with using the features directly or applying a softmax operation.

For the calculation of the ranking score, we apply a linear transformation to every pixel distribution, flattening the output to  $\mathbb{R}^{h \times w}$  and then an MLP with one hidden layer and ReLU activation. The final linear layer of the MLP generates a single scalar value representing the perception score.

It is important to note that the features given by segmentation are of considerably less dimensionality than traditional ResNet features and only capture the very specific information of which class is each pixel, which makes them significantly less expressive. Adding to that, since traditional CNN based approaches allow for finetuning, the amount of trainable parameters is also much smaller for this model than traditional models. Due to these two reasons the learning capacity of this model is much smaller, and therefore a significant performance drop is likely to happen.

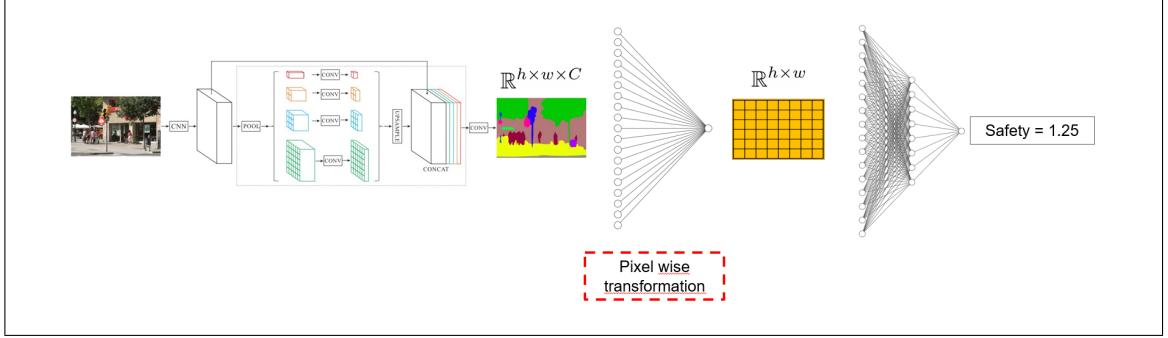


Figure 4.3. First model architecture

#### 4.2.2. SelfSegRank

With the intention of improving performance and explainability of SegRankBase model, we process the segmentation output with self attention mechanisms instead of a traditional MLP, since it has been proven to provide benefits in both aspects by previous research (Vaswani et al., 2017; Wiegreffe & Pinter, 2019; Cordonnier et al., 2019).

For our model we use the scaled dot product attention mechanism proposed by Vaswani et al. (2017). We abstain from using the full multi head attention mechanism that consists of the same operations but splitting the input in several “heads”. We do this because using multiple heads adds complexity to the interpretation of the attention outputs, since different heads may output inconsistent weights, as is mentioned in Clark et al. (2019) and J. Li, Tu, Yang, Lyu, and Zhang (2018) and also verified on this task by our own experiments.

The attention mechanisms receives three matrixes as input: the query  $Q$ , the key  $K$  and the value  $V$ . It calculates a matrix of attention weights over  $V$  based on  $Q$  and  $K$  and the final output is given by the product between  $V$  and the weights. A linear transformation is defined for each of the inputs with weights  $W_Q, W_K, W_V$  respectively, and another transformation  $W$  is applied to the final output (see Figure 4.4). Formally the attention

layer can be defined as:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (4.3)$$

$$\text{AttentionLayer}(Q, K, V) = \text{Attention}(QW_Q, KW_K, VW_V) \cdot W \quad (4.4)$$

where  $d_k$  is the embedding size of the key. For the particular case of self attention, the same input is used as query, key and value, so for our case we make  $Q = K = V = S'$  with  $S' \in \mathbb{R}^{(hw) \times C}$  and equal to the segmentation output flattened to one spatial dimension.

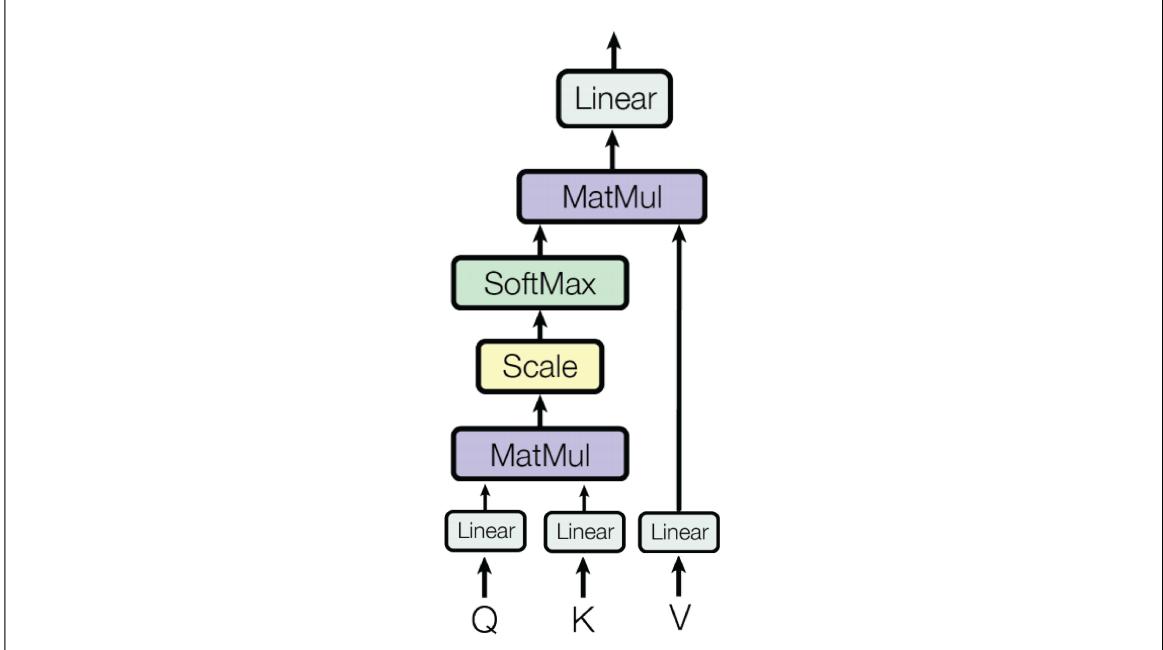


Figure 4.4. Attention layer operations. Adapted from Vaswani et al. (2017)

Similarly to the previous model, we apply a linear layer to calculate the ranking score from the attention output. We only use one layer instead of two in this model because the attention mechanism already has a large amount of parameters and a linear transformation of its own. In parallel the attention weights are also outputted by the model and are used for visualization. Figure 4.5 shows a diagram of the full architecture

As it can be seen on Figure 4.6, the attention weights keep the object shapes, allowing for a clear interpretation of which objects are significant to the output.

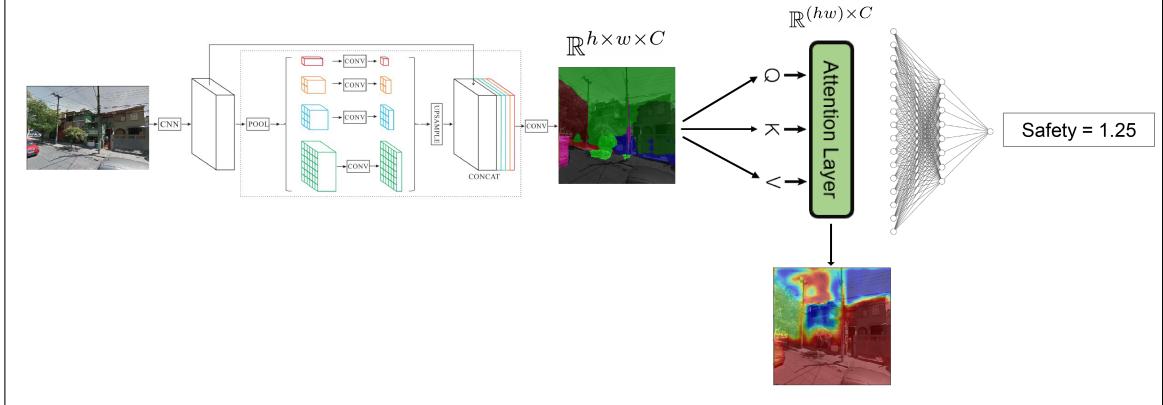


Figure 4.5. Segmentation and self attention network.

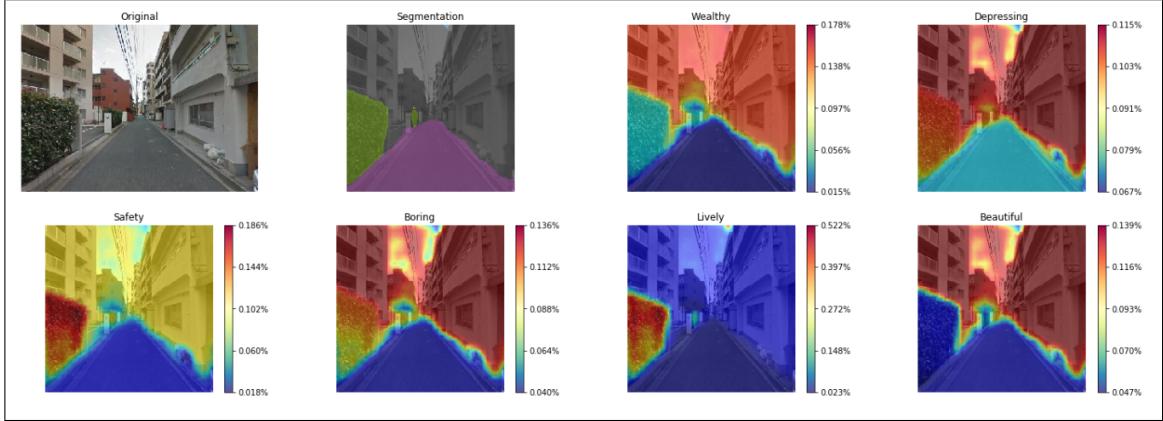


Figure 4.6. Example segmentation and self attention weights for all six attributes.

#### 4.2.3. AttentionSegRank

As was mentioned before, segmentation based features are considerably less expressive than traditional deep CNN features and cannot be finetuned, generating an important trade off between explainability and model performance. As a solution to that problem, we propose a mixed approach, that weights in both the image segmentation and the CNN features, in order to achieve both good performance and interpretability. To do that, we take advantage of the multiple inputs in the scaled dot product attention mechanism. The method consists of using the segmentation as key, and the ResNet features as both query and value (see Equations 4.3 and 4.4 ).

In practice,  $K$  and  $V$  must have the same spatial dimension for Equation 4.3 to be valid. We solve this problem by using layer *conv\_4f* of ResNet50 instead of *conv\_5c*, since it has a larger spatial dimensionality, and we add a transposed convolution layer (Noh, Hong, & Han, 2015) to do the final upsampling required to match the segmentation output dimension.

We do this because using the segmentation as key induces the attention weights to maintain a similar shape as the segmentation objects, keeping interpretability. To understand why this happens, see the  $QK^t$  product on Equation 4.3 that generates the weight matrix. A single element of the matrix (or a single attention weight) is given by:

$$a_{ij} = \sum_{l=1}^d q_{il} k_{jl} \quad (4.5)$$

With  $a_{ij}$  being the weight that the value feature  $v_j$  has on output feature  $i$ . Setting up  $K = S'$  and  $Q = F'$ , with  $S'$  and  $F'$  the flattened segmentation and ResNet features respectively:

$$a_{ij} = \sum_{l=1}^d f_{il} s_{jl} \quad (4.6)$$

Meaning that the weight of feature  $j$  on output feature  $i$  depends on which object is pixel  $j$ , resulting in an attention weight matrix that keeps the interpretability of the segmentation objects independently of the convolutional features.

### 4.3. Additional components

#### 4.3.1. Loss function

The loss function for this task must account for the pairwise structure of the dataset, and should represent the cost of breaking restrictions given by Equations 4.1 and 4.2. For 4.1 we use a hinge loss similar to the one proposed by Dubey et al. (2016):

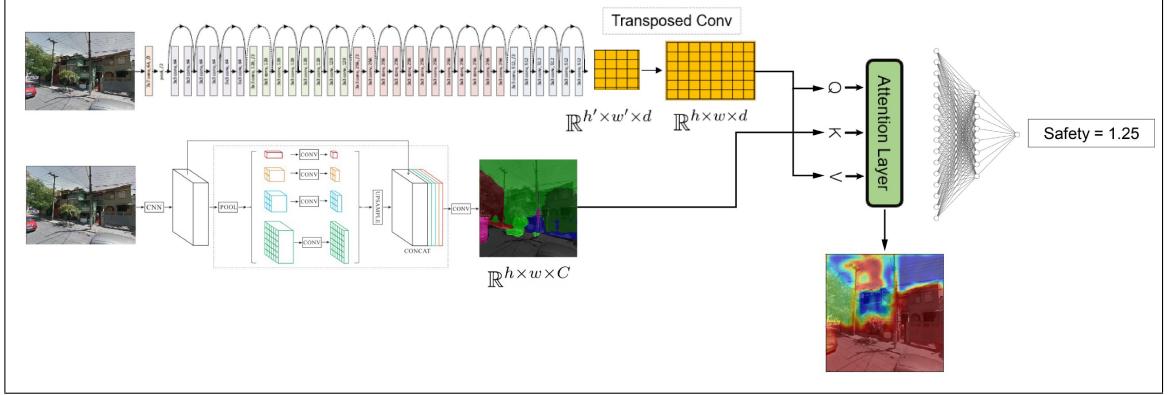


Figure 4.7. Segmentation as key network.

$$L_r(x_1, x_2, y | \Theta) = \max(0, -y(f_\Theta(x_1) - f_\Theta(x_2)) + m_r) \quad (4.7)$$

Where  $f_\Theta$  and  $\Theta$  represent the network and its parameters respectively, and  $m_r$  is an hyperparameter. This loss component makes it so that the model learns to assign a higher score to the image winner of the vote. Based on the work by Doughty, Damen, and Mayol-Cuevas (2018) we also add a second component so that tied votes can be used for training. According with Equation 4.2 we define:

$$L_t(x_1, x_2 | \Theta) = \max(0, |f_\Theta(x_1) - f_\Theta(x_2)| - m_t) \quad (4.8)$$

Where  $m_t$  is also an hyperparameter. Finally, the complete loss function is defined as:

$$L(x_1, x_2, y | \Theta) = \begin{cases} L_r(x_1, x_2, y) & \text{if } y \in \{-1, 1\} \\ L_t(x_1, x_2) & \text{if } y = 0 \end{cases} \quad (4.9)$$

In practice we take the mean loss over the batch examples and we set  $m_r = m_t = 1$

### 4.3.2. Semantic Dropout

It's important to note that the semantic segmentation model trained on cityscapes, presents an unavoidable drop in segmentation performance when applied on PlacePulse due to domain shift. Errors in the segmentation can produce significant problems in the final perception quantification and can also cause confusing attention heatmaps due to errors in the object edges.

In practice we identified a tendency for the models to have attention weights highly biased towards specific segmentation classes, which is highly undesirable both for explainability and model generalization.

We solve these problems by implementing what we call Semantic Dropout, similar to traditional Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), but instead of dropping a single neuron with probability  $p$ , Semantic Dropout sets the probabilities of an entire segmentation class to 0, inducing artificial errors during training and preventing the network from becoming too sensitive to segmentation errors while also reducing the bias in attention weights. This effect can be seen on Figure 4.8, which shows how the attention weights are less biased towards one class and therefore less affected by the errors in segmentation. Mathematically this technique is equivalent to the spatial dropout proposed in Tompson, Goroshin, Jain, LeCun, and Bregler (2015), but applied to segmentation probabilities instead to convolutional kernels.

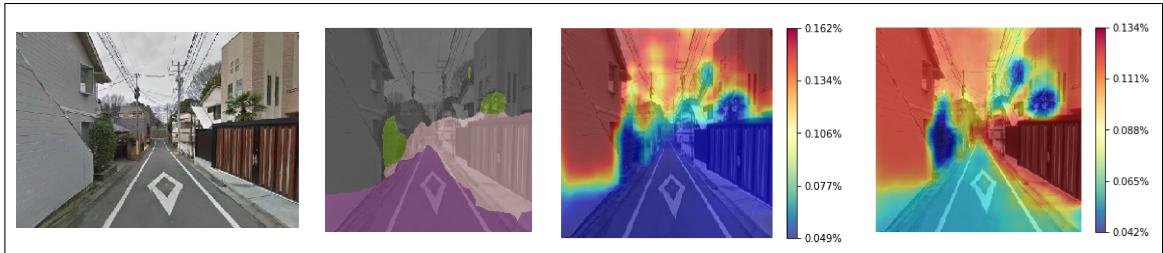


Figure 4.8. Effect of Semantic Dropout. From left to right: original image, semantic segmentation and the outputs of the self attention model trained, without and with semantic dropout.

## 4.4. Baselines

With the purpose of making and ablation study, we also train two baseline models based on the architecture proposed by Dubey et al. (2016), designed for measuring the effect of the segmentation and attention mechanisms in both performance and explainability of the models. We base these models on the ResNet50 CNN (He et al., 2015), as is the defacto approach for computer vision problems. We abstain from using larger versions of ResNet due to significant overfitting issues.

### 4.4.1. ResNet50 + MLP

The first baseline consists of a standard finetuned ResNet50 with a two layer MLP. This model doesn't provide any sort of out of the box explainability and therefore is useful to measure how segmentation affects performance. Unlike its segmentation based sibling, dropout and L2 regularization are necessary for training this model, due to the significantly larger amount of trainable parameters that come from finetuning the CNN.

### 4.4.2. ResNet50 + Self Attention layer + MLP

A baseline on explainability is also important, since improving it is the key contribution of this work. For that we use a similar architecture to attention-based explainability models from the literature (Y. Zhang et al., 2019; Cordonnier et al., 2019; Bello et al., 2019), consisting on combining a finetuned CNN with self attention layers.

We take the output of ResNet50's *conv\_5c* layers and give it to the attention layer defined in Section 4.2.2 and then to a two layer MLP for calculation of the final score.

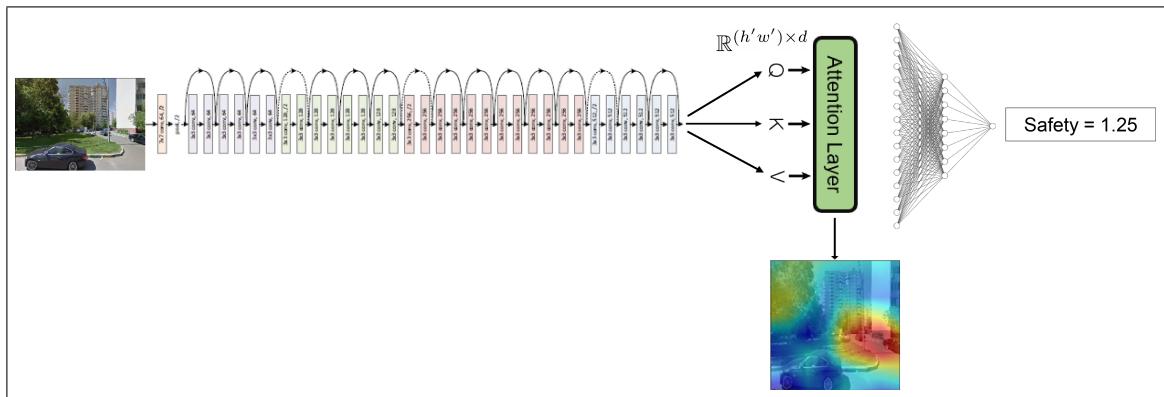


Figure 4.9. ResNet and self attention network. ResNet diagram extracted from He et al. (2015).

## 5. METHODOLOGY

This chapter shows the practical details of implementation and training.

### 5.1. Models and training

All of our models are implemented using the Pytorch library (Paszke et al., 2019) version 1.2.0. We use the implementation and pretrained weights of ResNet available on the Torchvision library (Marcel & Rodriguez, 2010). We train our own PSPNet based on the implementation by Huang, Wei, Wang, and Liu (2019). All models are trained using a single 12 Gb Nvidia Geforce-GTX 1080 Ti GPU except for the mixed model, which is trained on a 24 Gb Nvidia Titan RTX.

For training we make a 75%/25% train/validation splits of the dataset for each attribute. We keep the splits fixed for all models, so they all see and are evaluated on the same data. All models are trained for 40 epochs and we keep the model with the best validation accuracy on epoch end.

Table 5.1. Hyper parameters and configurations for each model.

Parameter/Model	ResNet50	ResnetAttn	SegRank	SelfSegRank	AttentionSegRank
Batch Size	32	32	32	32	32
Learning Rate	$10^{-4}$	$10^{-4}$	$10^{-4}$	$10^{-4}$	$10^{-4}$
Opt. Algorithm	SGD	SGD	Adam	Adam	Adam
Finetuning	Yes	Yes	No	No	Yes
Dropout	0.3	0.3	0	0	0.1
Semantic Dropout	N/A	N/A	0	0.1	0.1
Weight Decay	$10^{-5}$	$10^{-5}$	0	0	0

Baselines are trained with SGD with a momentum of 0.9 (Rumelhart, Hinton, & Williams, 1986) as it provided better results empirically. For segmentation based models we train with Adam (Kingma & Ba, 2014) and we set  $\epsilon$ ,  $\beta_1$  and  $\beta_2$  to  $10^{-9}$ , 0.9 and 0.98 respectively. We use semantic dropout on both models that have segmentation and attention, and add an equivalent regular dropout layer to the ResNetAttn Baseline for fair comparison. Weight decay and traditional dropout are used for all baseline models that finetune ResNet weights. See Table 5.1 for details on the training hyperparameters.

## 5.2. Visualization

For visualization we generate both segmentation and attention images, in this section we will explain how we generate the attention visualizations. See Appendix A.1 for segmentation.

Recall that our attention matrix  $A$  is of dimensions  $(hw) \times (hw)$ , meaning that we have for each pixel, the attention that this pixel gives to every other pixel. To reduce the matrix to a single value per pixel  $p$ , we average the attention given by every pixel to  $p$ . This can be done by simply taking the column wise mean of  $A$ , which results in a vector with  $hw$  components. Finally, we reshape this vector to the original input size, obtaining an attention map  $A' \in \mathbb{R}_{[0,1]}^{h \times w}$ .

Since in practice,  $h$  and  $w$  are too small to produce a good quality visualization we resize  $A'$  with bilinear interpolation to the standard size of  $224 \times 224$ . Finally each image is min max normalized to be in the  $[0, 255]$  interval in order to apply the color gradient that generates the final heatmap. For sample results see section 6.2.

The heatmaps serve as an additional model interpretation tool on a per instance level. For example, showing the visualization for all attributes on a single image, shows how the models pay attention to different parts of the image for different attributes.

## 6. RESULTS

This chapter shows the main results obtained. On Sections 6.1.1 and 6.1.2 we present the quantitative performance and training results. Section 6.2 explains how the different visualizations are generated, including examples for all the models.

### 6.1. Quantitative results

#### 6.1.1. Model performance

Even though the objective of this research is to learn a ranking (or regression) to quantify the urban perception, exact labels for this are not available, so we have to measure model performance based on the Place Pulse votes, which as was mentioned on section 4.1, has considerable issues. We use as performance measure the equivalent to classification accuracy, considering which image won the vote as the target label. In other words, we evaluate the percentage of restrictions (see 4.1) that are satisfied by the model. We do this separately for each attribute in its corresponding validation set and the final accuracy value for each model is calculated as the mean accuracy through all attributes.

Both ResNet based baseline models achieve an accuracy of ~66% and as it was expected, replacing the more expressive CNN features for semantic segmentation, caused a significant performance drop, falling to 60.62% for SegRank and 61.44% for SelfSegRank. The performance is then mostly recovered when combining both approaches, reaching 64.41% on the AttentionSegRank architecture. See Table 6.1, for the exact accuracy values.

Attribute/Model	Resnet50	ResnetAttn	SegRank	SelfSegRank	AttnSegRank
Wealthy	68.45%	68.60%	60.99%	61.89%	66.27%
Depressing	65.87%	66.02%	61.73%	62.30%	64.33%
Safety	65.41%	65.46%	59.08%	60.24%	64.04%
Lively	65.93%	64.78%	58.96%	60.29%	63.35%
Boring	62.28%	62.20%	58.01%	58.51%	61.06%
Beautiful	70.02%	69.60%	64.93%	65.43%	67.99%
Average	<b>66.33%</b>	66.11%	60.62%	61.44%	64.51%

Table 6.1. Accuracy values for each model on all attributes.

### 6.1.2. Training behavior

Models trained on the place pulse dataset are very prone to overfitting. We believe this is due to it having a very large amount of votes in comparison to the amount of available images, and because the task is very hard to generalize given the high amount of noise that the dataset has from how it was collected. This can be seen clearly on figures 6.1 and 6.2. Both baselines models present considerable overfitting, showing accuracy differences between seen and unseen data of up to 25%, and ceasing to improve on the validation set after one or two training epochs.

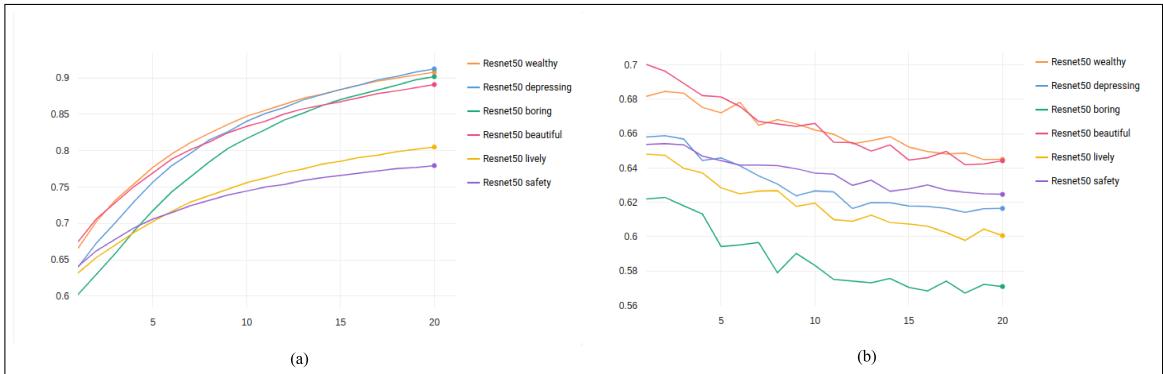


Figure 6.1. ResNet50 baseline accuracy vs epoch learning curves on training (a) and validation (b).

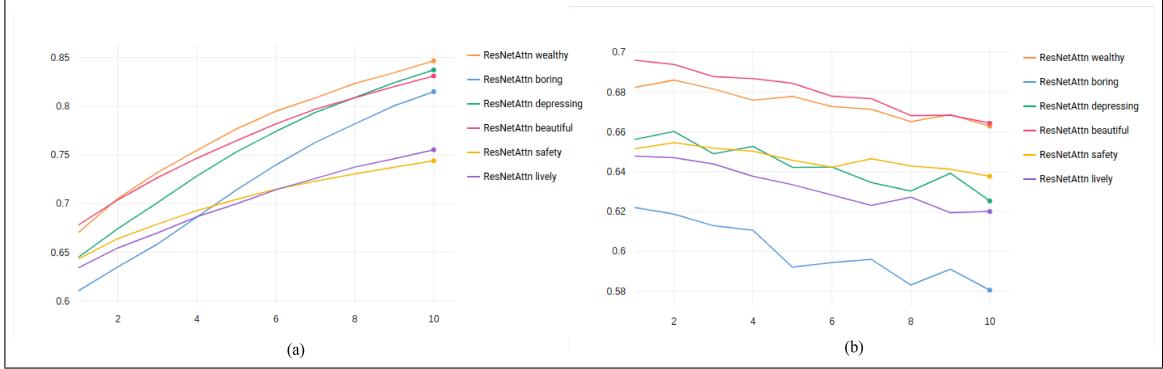


Figure 6.2. ResNetAttn baseline accuracy vs epoch learning curves on training (a) and validation (b).

Replacing the CNN features for semantic segmentation generates a considerable change in training behavior, with the reduced expressiveness of the segmentation overfitting completely disappears, which translates to a drop of around 20% to 30% accuracy in training, but of only 6% on validation.

The basic SegRank architecture still reaches convergence after one or two epochs. Adding the self attention layer makes convergence slightly slower allowing the model reach a higher validation accuracy.

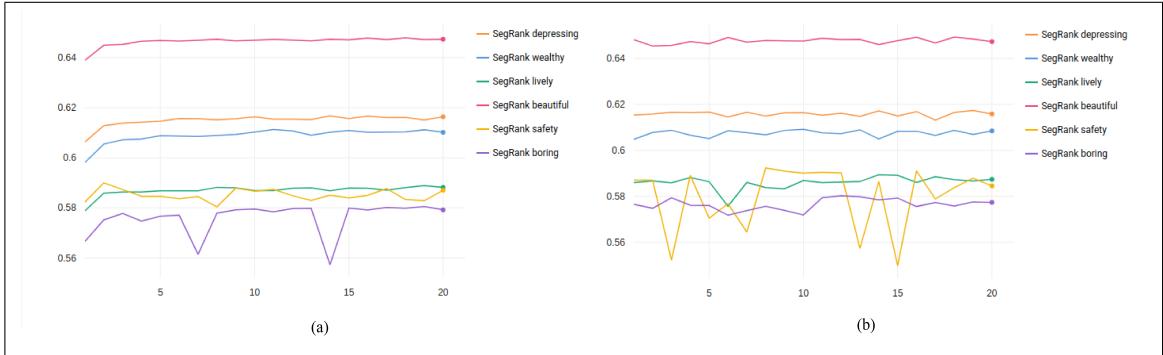


Figure 6.3. SegRank accuracy vs epoch learning curves on training (a) and validation (b).

Interestingly, similar but reduced behavior happens with the AttentionSegRank model, dropping around 10% to 20% accuracy on training and 2% on validation.

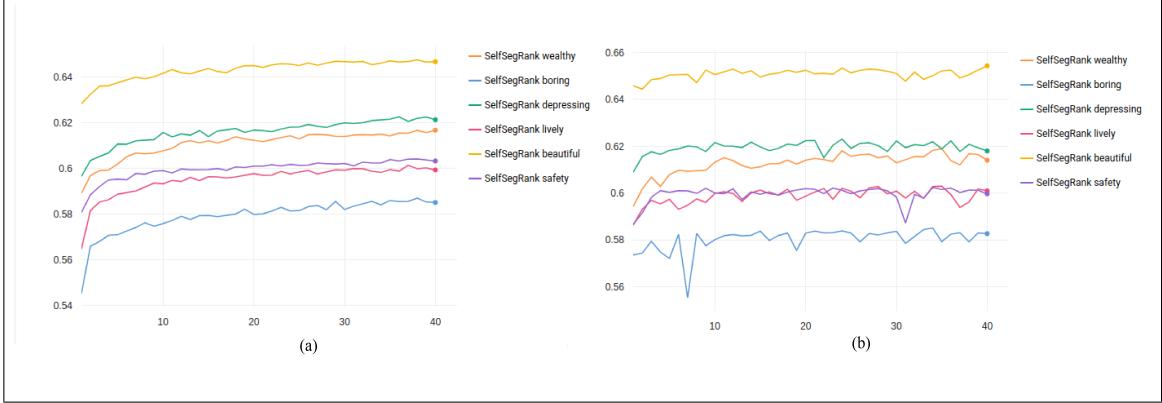


Figure 6.4. SelfSegRank accuracy vs epoch learning curves on training (a) and validation (b).

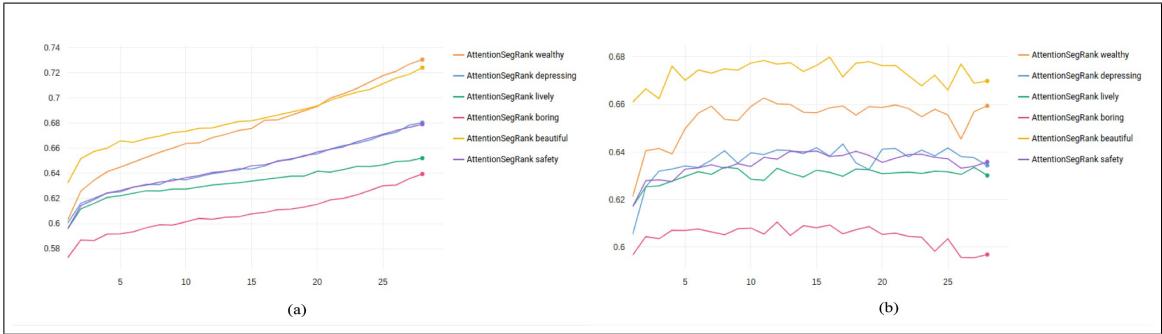


Figure 6.5. AttentionSegRank accuracy vs epoch learning curves on training (a) and validation (b).

As it can be seen on the results of all models, the learning process is consistent throughout the different attributes. The accuracy of the different attributes is also consistent across the different models, with boring and beautiful being the hardest and easiest tasks to learn respectively on all models.

## 6.2. Visualization results

As was detailed in Section 5.2, we derive heatmaps from the attention weights, which serve as an additional model interpretation tool on a per instance level. For example, showing the visualization for all attributes on a single image, shows how the models pay

attention to different parts of the image for different attributes. See figures 4.6, 6.6 for examples.

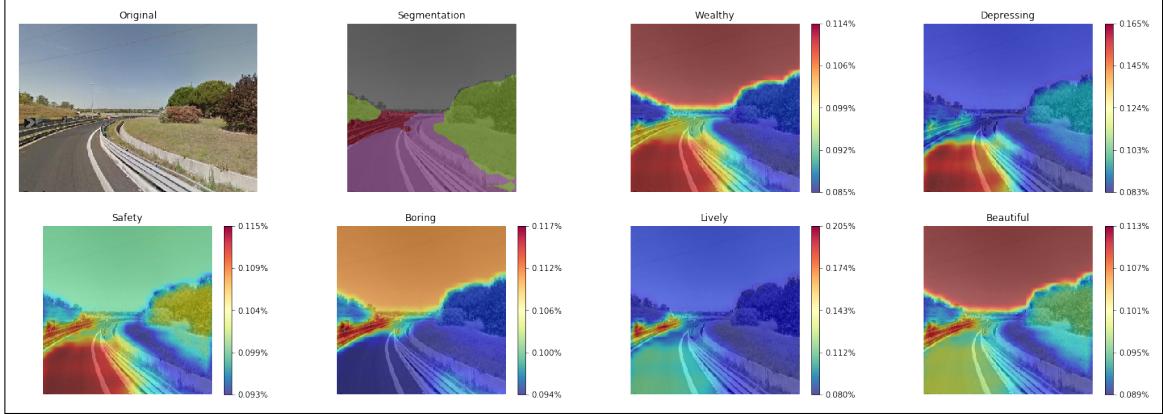


Figure 6.6. Attention heatmaps for the AttentionSegRank architecture on all attributes.

Combining the segmentation with the heatmap allows to see how the attention is distributed through the different segmentation classes, both qualitatively and quantitatively, as can be seen in Figure 6.7.

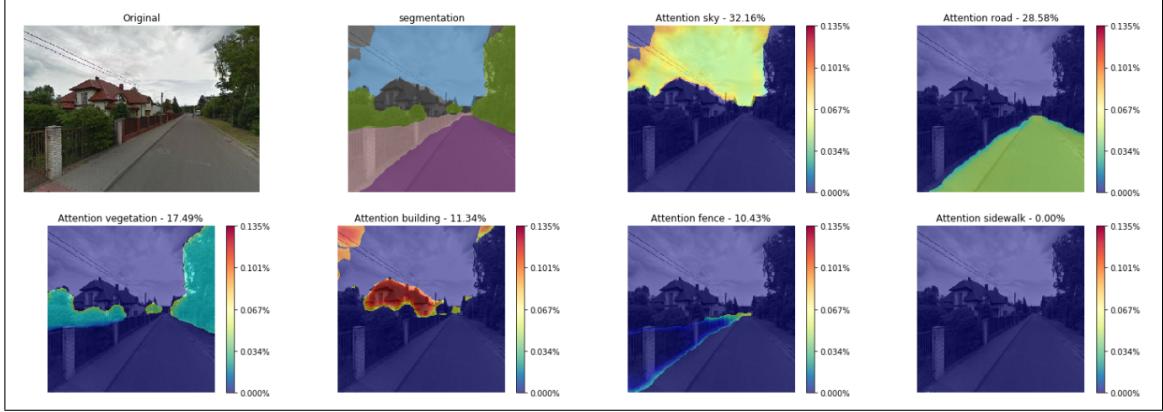


Figure 6.7. Segmentation class separated attention heatmaps for the AttentionSegRank architecture on the boring attribute. Only classes present on the segmentation are shown. Percentages indicate the amount of attention given to that class.

It is important to note how the three attention models behave differently for the same image, which can be seen by plotting the heatmaps side by side, as it is shown in Figure 6.8

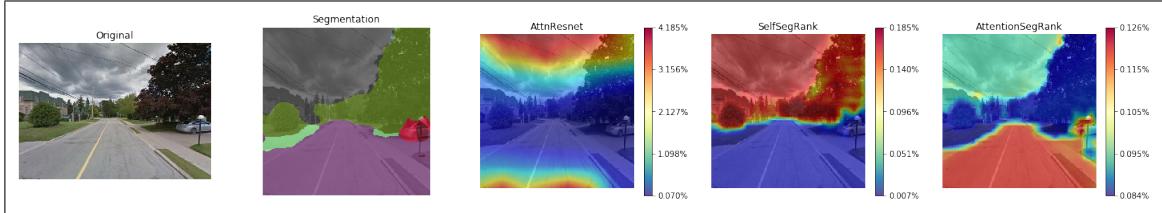


Figure 6.8. Example of attention heatmaps for the three models on the safety attribute. Using segmentation adds significant improvement to the heatmap shapes.

Finally by gathering up the three model outputs a very informative visualization of the urban perception of an image can be built. See figures 6.9 and 6.10 which show ranked samples for each of the attributes along with the segmentations and attention heatmaps.

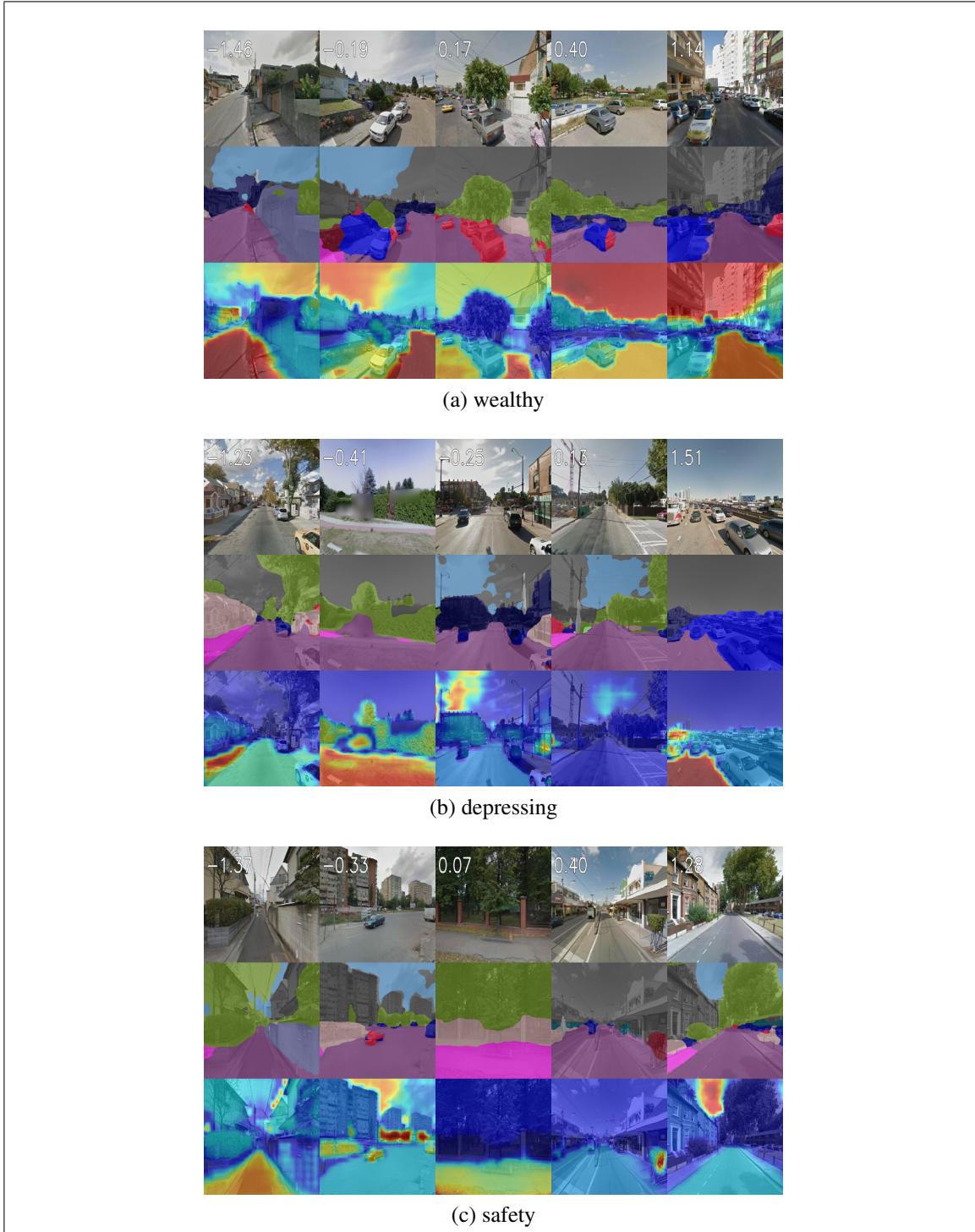
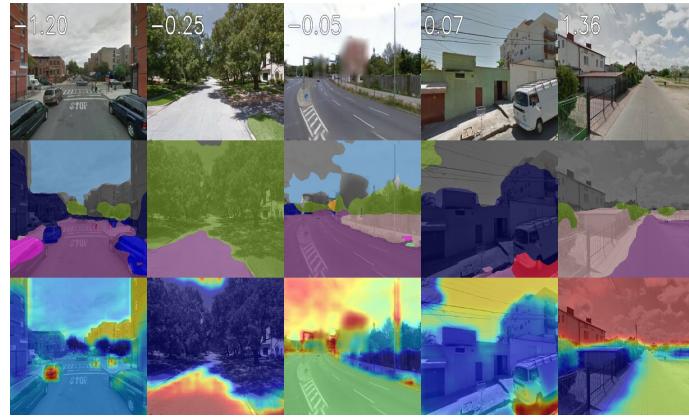
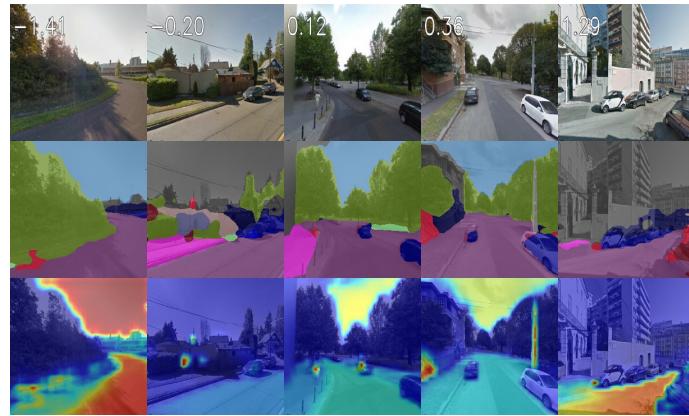


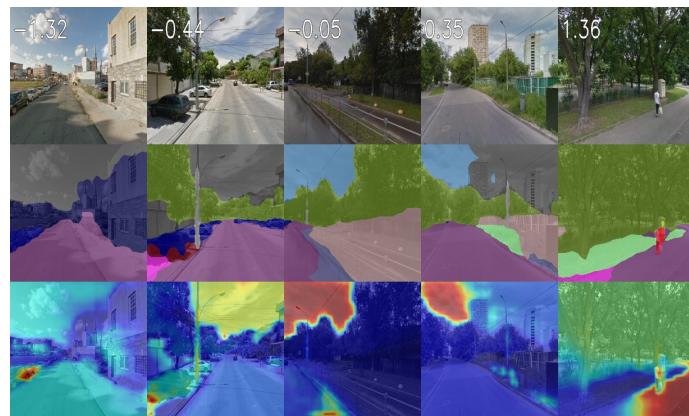
Figure 6.9. Result samples for wealthy, depressing and safety, ordered from lower to higher score. The samples are taken from the validation split of each attribute.



(a) boring



(b) lively



(c) beautiful

Figure 6.10. Result samples for boring, lively and beautiful, ordered from lower to higher score. The samples are taken from the validation split of each attribute.

## 7. DISCUSSION

In this chapter we will do further analyses of the model results with particular emphasis on the AttentionSegRank architecture due to its better performance and explainability. On Section we will discuss the effects of using semantic segmentation on neural network training. Sections 7.2 and 7.3 will show a quantitative analysis between segmentation, attention, and the perception quantification. Section four shows the effect of combining image features with segmentation. And finally, on Section five we will analyze the implications of this method on model explainability.

### 7.1. Effect of semantic segmentation on learning

As was already mentioned on section 6.1.2, adding a fixed segmentator to the neural network architecture resulted in a reduction of performance along with a considerable reduction of overfitting. The behavior was expected when fully replacing the CNN features, due to the reduced expressiveness of the segmentation and the lack of finetuning, but unexpectedly, although it is reduced, this behavior persists when combining the fixed segmentation with the finetuned ResNet50 through the attention layer. We conclude from this that restricting the attention weights to the shapes and classes given by the segmentation has a regularizing effect on learning, reducing the model capacity even when the amount of trainable weights is maintained.

In the case of the PlacePulse dataset this is not a problem since all traditional deep models suffer of significant overfitting. It remains an interesting research question if these behaviors will transfer to other tasks and datasets.

### 7.2. Segmentation as an attention subject.

By combining the outputs of the segmentation and the attention, we can infer if a certain class was important for the quantification of perception of a single image. We do

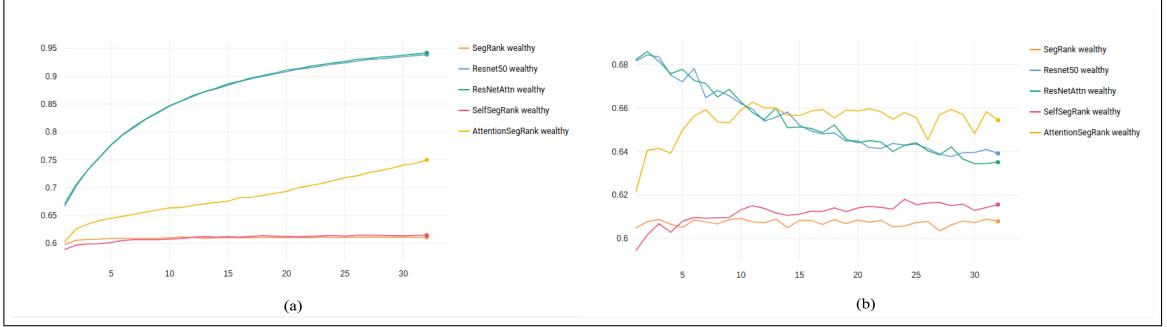


Figure 7.1. Wealthy accuracy vs epoch learning curves on training (a) and validation (b).

this by computing the ratio between the percentage of pixels belonging to a single class and the percentage of attention placed by the model on those pixels:

$$\text{significance ratio} = \frac{\% \text{ attention}}{\% \text{ segmentation}} \quad (7.1)$$

Using that, we define that a class is significant in an image when significance ratio  $\geq 1$ , meaning that the total attention received by the pixels belonging to that class is larger than the amount they would get if the attention was distributed uniformly over the whole image.

We calculate the significance of all the CityScapes classes on all the images of the validation split for each of the perception attributes. We show the results for the most significant classes on figures 7.2 and 7.3, which show the percentage of images where each class is considered significant, over the whole split or filtered by the images where the class is present respectively. For the full results see appendix C .

The most notable insight from this results is that there are considerable differences on the same class but for different attributes, for example buildings are the most significant for the boring attribute but is one of the least for the depressing attribute. Vegetation is significant for beautiful and depressive but not for wealthy. Another insight is that some classes are on average considerably more significant than others, we believe this happens because they are much more common than others in the dataset and the networks learn

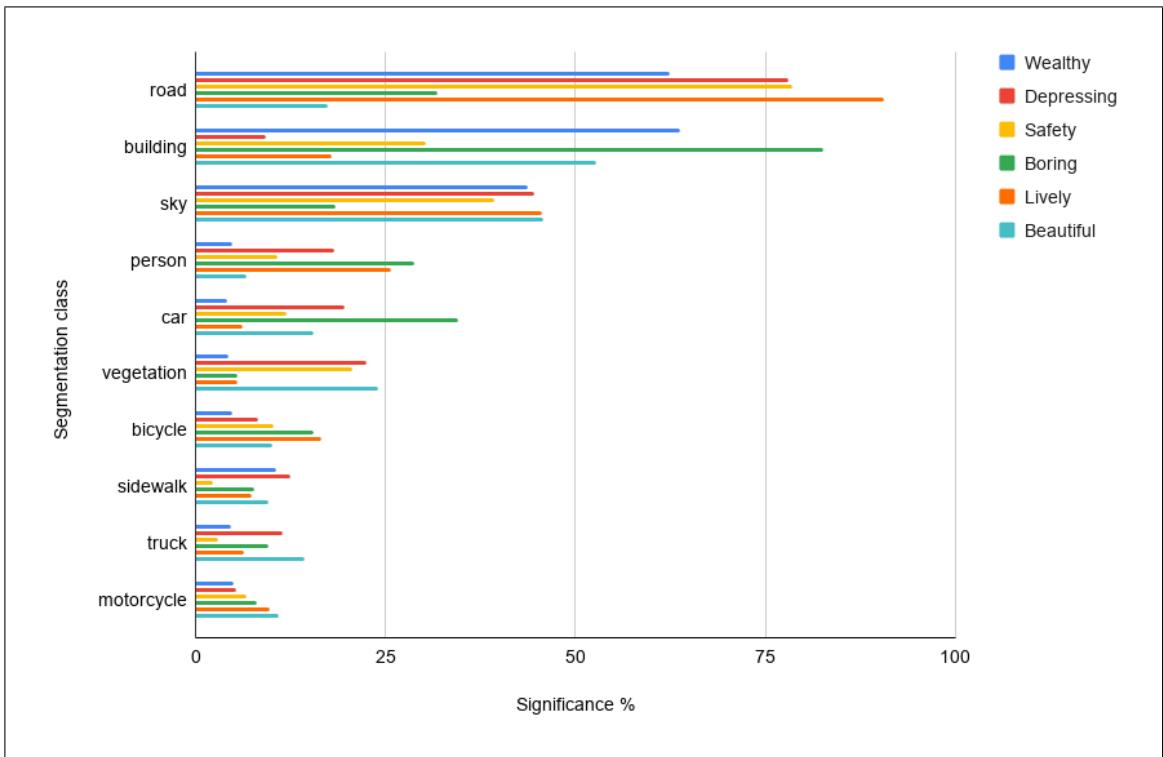


Figure 7.2. Percentage of times a segmentation class is considered significant over the whole dataset, for the 10 most significant classes on average.

to consider them more often. For the exact distribution of the segmentation classes see appendix A.2.

Even though the set of classes used by F. Zhang et al. (2018) is not exactly the same as the one we use in our study, the most significant classes for our model are consistent with the ones presented in their regression model. This is important because it shows we are approaching the interpretability of regression model with a considerably larger neural network.

These results support that attention weights are a good way to augment the explainability of the models since they show that the same architecture learns to attend to different things when quantifying different attributes.

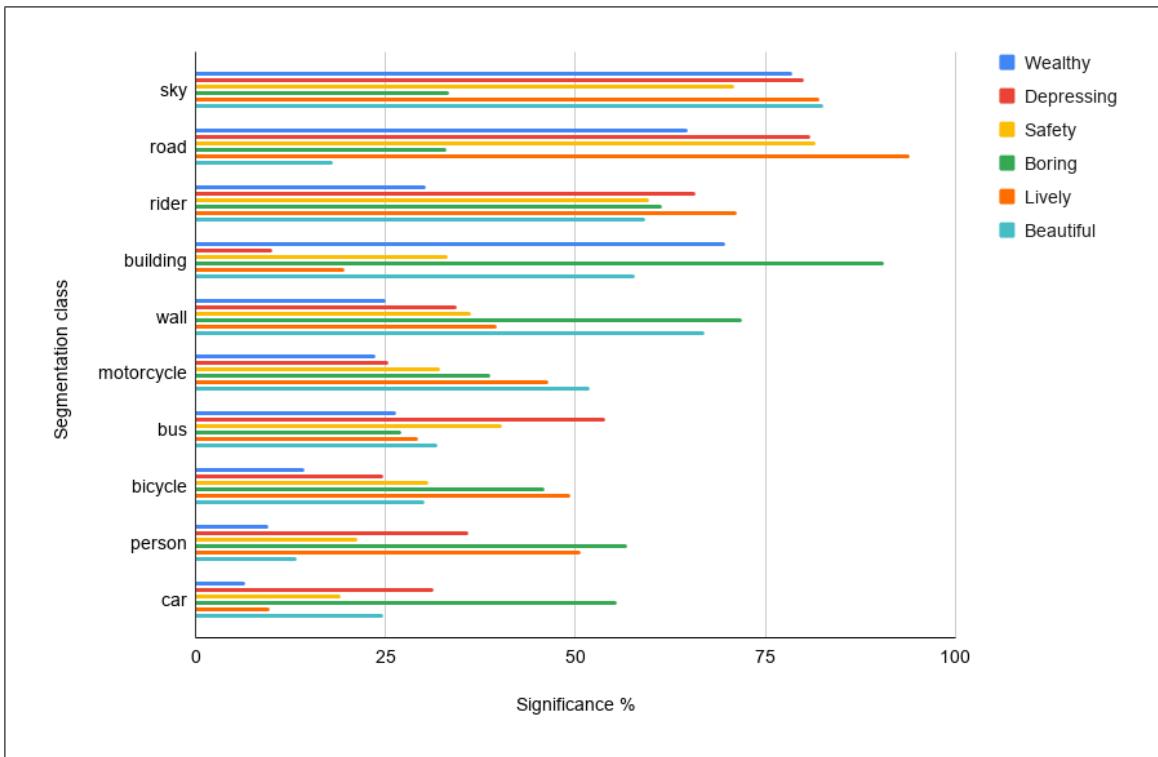


Figure 7.3. Percentage of times a segmentation class is considered significant when it is present in an image, for the 10 most significant classes on average.

We can also visualize which classes are significant for individual examples, allowing for a simple interpretation of results on a per instance basis. See Figure 7.4 in which the road and sky classes were determined as significant by the models attention.

### 7.3. Relationship between urban perception and semantic segmentation

Similarly to previous work (Rossetti et al., 2019; F. Zhang et al., 2018) we analyze how the segmentation correlates with the quantified perception scores. For that, we count the percentage of pixels of each segmentation class on all the images of the validation split for each of the perception attributes and measure correlation with the perception quantification. We present these results for both segmentation and attention percentages on Figure 7.3.

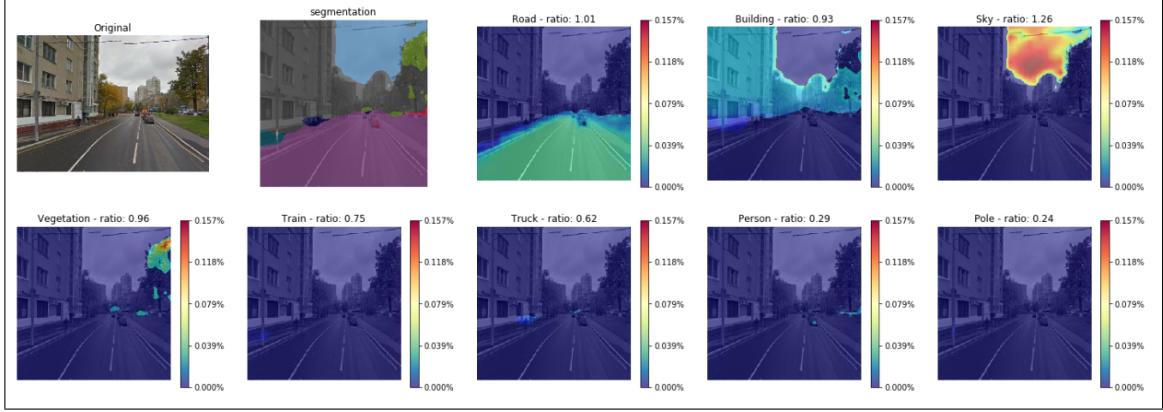


Figure 7.4. Sample visualization including class significance for the safety attribute.

In all classes correlation is not extremely high, we believe that to be a good result since no class should have such a determinant inherence on the perception of an attribute as to reach a correlation close to 1 or -1. Attention and segmentation behave similarly, having opposite sign only on very rare cases, (e.g: road for the lively attribute) and differing only on the magnitude of the correlation most of the time. Comparing with the significance results shown on Figure 7.2, the most significant classes on the dataset are also consistently on the top or bottom of the correlation charts, so the high significance comes hand in hand with high correlation more often than not.

As is expected, the model’s outputs correlate with the classes differently depending on the attribute, making sense with how the attributes themselves are correlated, for example vegetation is very positive for beauty but very negative for depressiveness.

We also compute correlation between segmentation and score when the respective class is considered significant by the attention weights. The results are mostly maintained but with a slight tendency to have larger correlation magnitude. Figure 7.6 shows an example comparing the correlation for all the images and filtered for significance. Most classes get an increase in magnitude while keeping the sign fixed. Sign changes can happen occasionally for classes with very low correlation magnitude ( $< 0.1$ ).

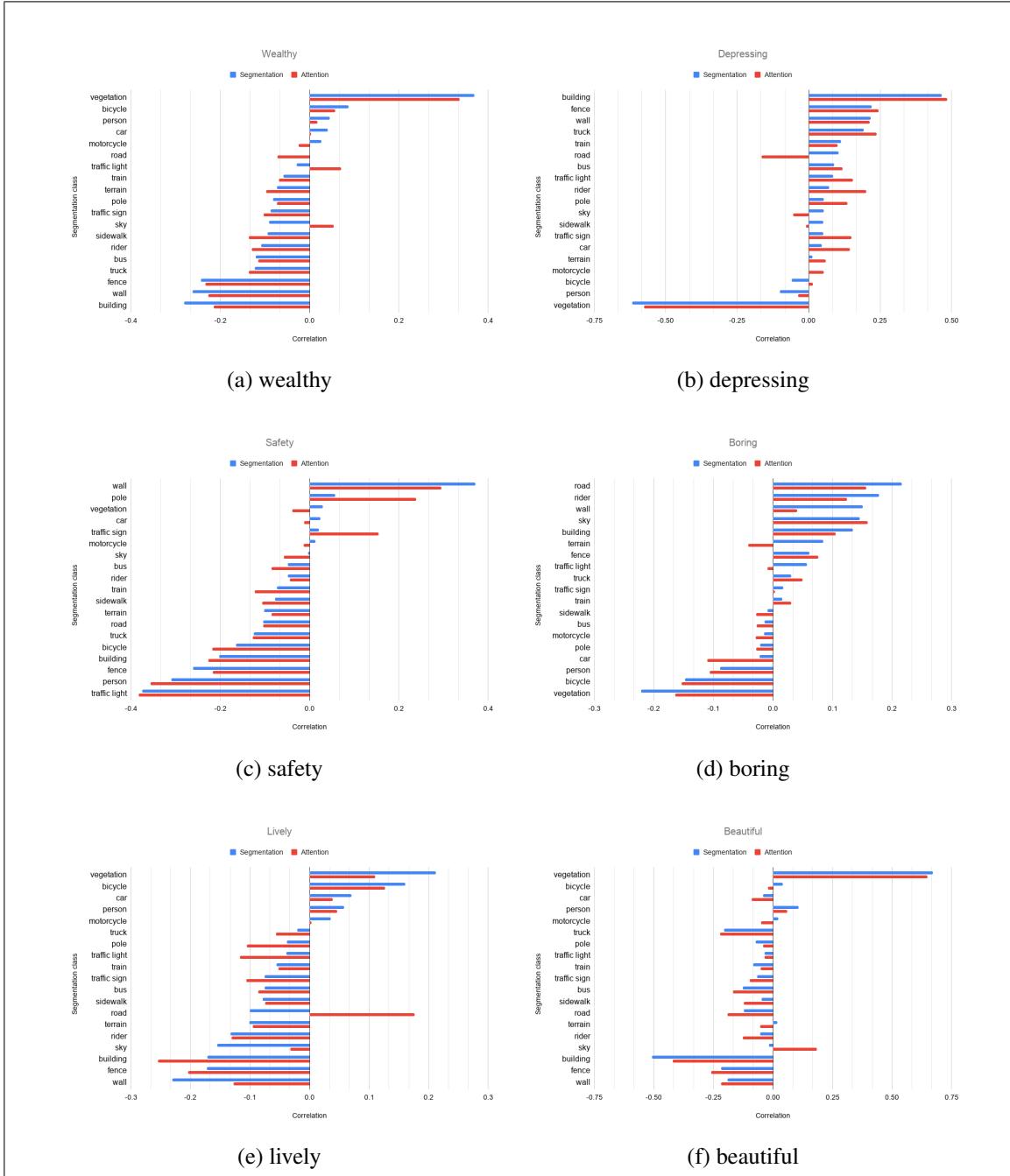


Figure 7.5. Correlation between segmentation and attention percentages for each class and quantified scores.

Even though some classes have high (or low) correlation is important to note that in no case their sole values are enough to determine the final output score, even when filtering

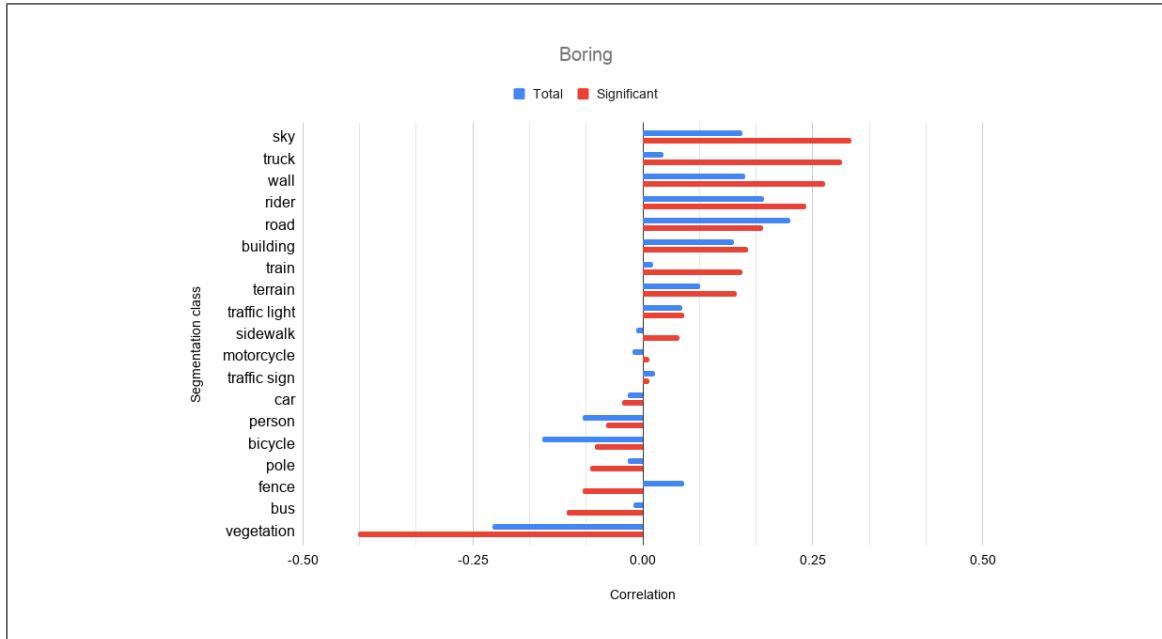


Figure 7.6. Correlation between score and segmentation percentage. Comparison between the whole dataset and filtering each class by significance. Example for the boring attribute.

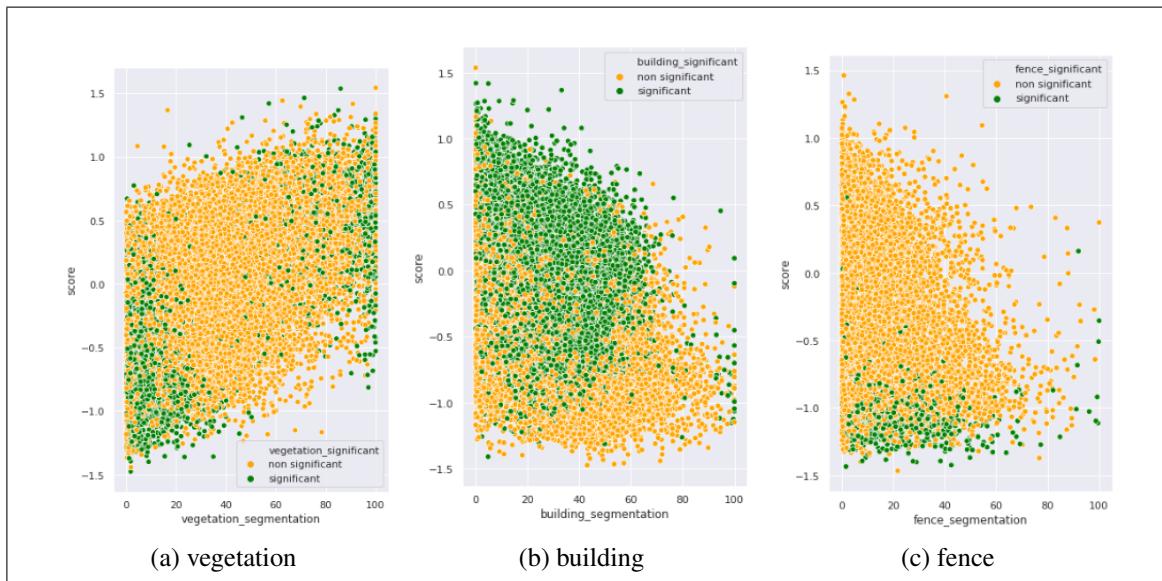


Figure 7.7. Scatter plots for the 3 classes with highest correlation for the beautiful attribute. Colors denote if that class was significant for that image.

by significance. Other objects on the images and ResNet features are also necessary to get a more precise quantification of the perception. See Figure 7.7 where even though the correlation can be seen clearly, for a single segmentation percentage the scores fall over a large interval.

As we mentioned above, F. Zhang et al. (2018) used different segmentation classes in their study so an exact comparison of results is not possible. However the correlations we obtained are mostly consistent in sign and in order of magnitude with the beta coefficients they obtained (see figures 7.3 and 2.2).

Some interesting insights can be drawn from the correlation results:

- Buildings have a general "negative" connotation, making images more depressing and boring, and less wealthy, safe, lively and beautiful.
- Vegetation is the opposite, making images more wealthy, lively and beautiful and less depressing and boring.
- Unlike (Rossetti et al., 2019) we find that fences have a general "negative" connotation, similar to buildings.
- People make images feel less safe, but also less boring.
- Bicycles, which are rarely present in the dataset, are very discriminative for many attributes.
- Even though it has high attention weights on a large percentage of the images, the sky doesn't have a clear tendency on most attributes, excepting for boring.

#### **7.4. Effect of combining features and segmentation.**

As a way of ablation study we evaluate the same metrics with the more basic models. By analyzing SelfSegRank and AttentionSegRank we can see the impact the ResNet features have on the model behavior besides the performance improvement. Using only segmentation results in a larger bias in significance towards the classes more common in the dataset, with the least present reaching 0% significance for some attributes. Following

that tendency, the correlation order of the classes is mostly kept, but with an increase in the correlation magnitude, that is an expected behavior since this model depends on segmentation as the sole input. An example of this difference can be seen on Figure 7.8 or by comparing Figure 7.9 and Figure 7.5a.

Another interesting behavior of the segmentation only architectures is that when a single segmentation class takes a large portion of an image, the model output tends to converge to a middle point value. This may be due to the lack of information the model has when the inputs consists solely of a large blob of a single class with no additional features. The fact that this behavior is not replicated by the AttentionSegRank model, supports that adding the traditional image features to the segmentation is a better approach not only for performance but for explainability as well.

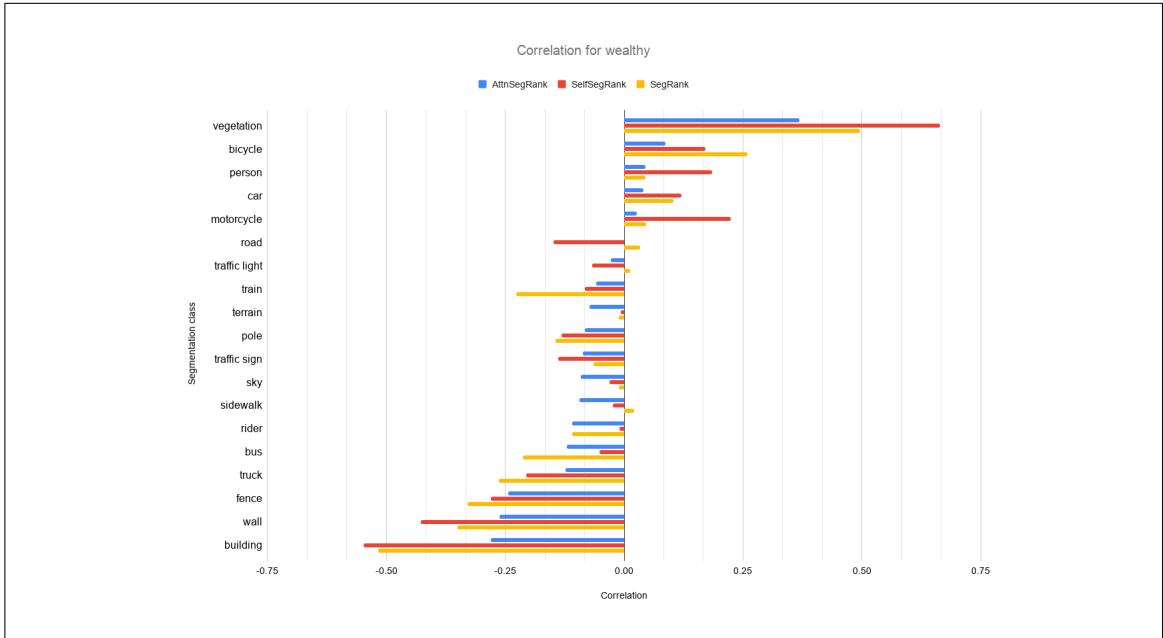


Figure 7.8. Segmentation and score correlation comparison for the three architectures on the wealthy attribute.

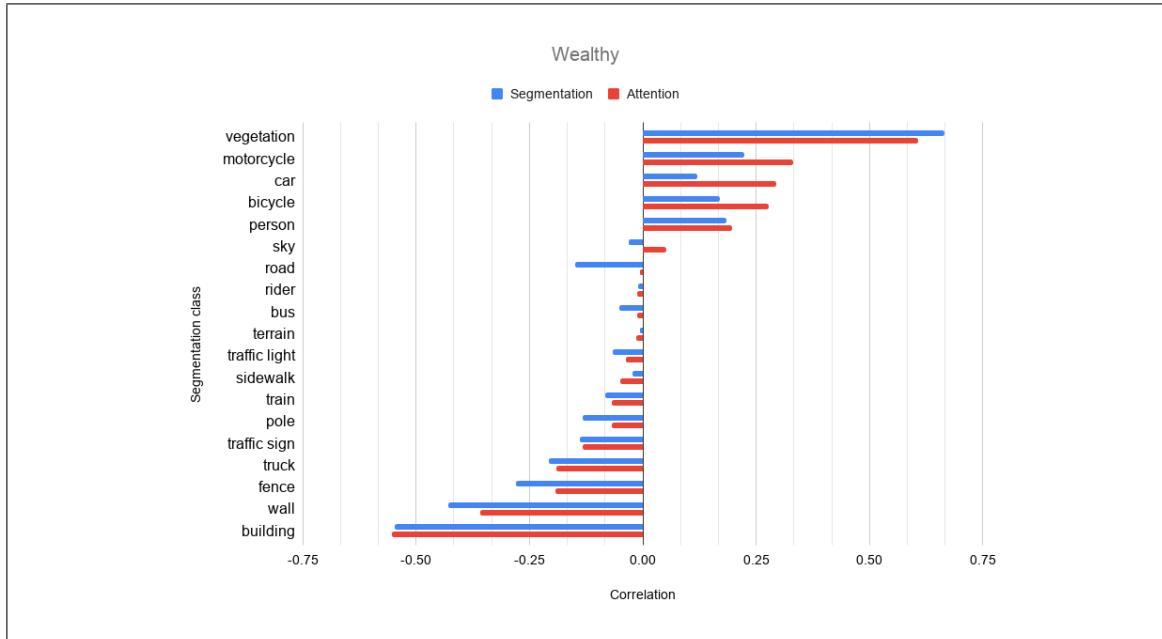


Figure 7.9. Correlation between score and segmentation/attention percentages on the SelfSegRank model. Example for the SelfSegRank architecture on the wealthy attribute.

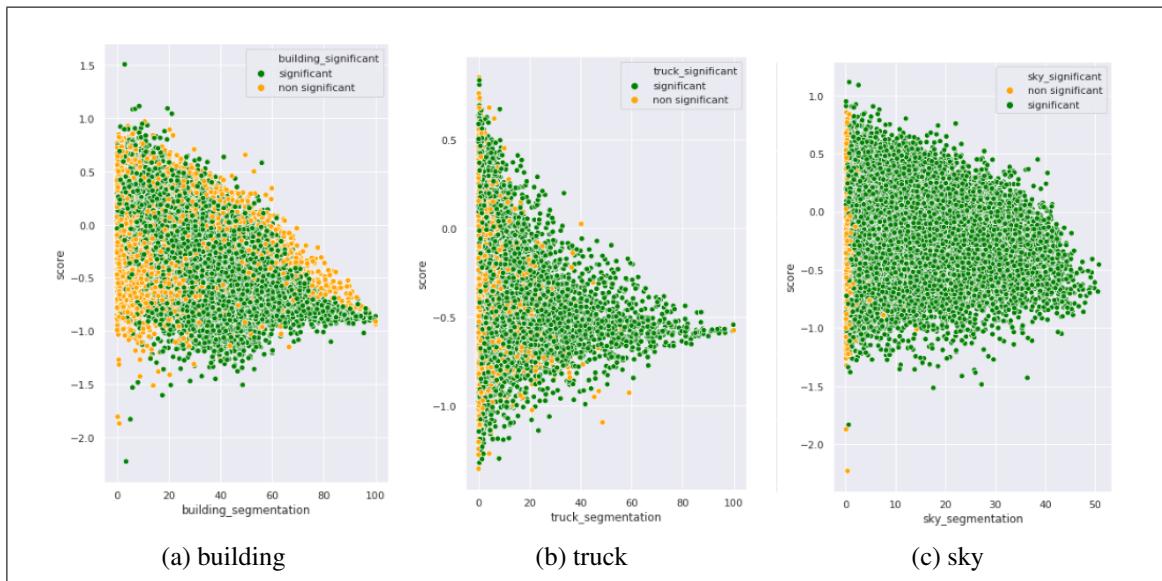


Figure 7.10. Scatter plots showing examples of converging attributes on the SelfSegRank model. All plots made for the wealthy attribute.

## 7.5. Effect of attention over semantic segmentation on model explainability.

The main objective of this work is to improve model explainability. The tool we present towards that objective is the use of semantic segmentation to compute attention weights. This approach generates considerably better attention weights from a semantic standpoint, since the shapes generated have a very high consistency with those given by the semantic segmentation, which are by definition human understandable.

Models in the computer vision literature that use attention weights to enhance explainability have the constant problem of region boundaries not being clearly delimited, often having shapes that do not resemble any human known object or figure in the image (Y. Zhang et al., 2019; Z. Zhang et al., 2020; Johnston & Carneiro, 2020; Carion et al., 2020). Figure 6.8 shows an example of the advantage of our method over the baseline, which consists of a traditional attention mechanism.

Another benefit of segmentation based attention weights is, since weights have a human understandable shape, we can also analyze how they behave visually in an aggregated way, we do that by calculating the mean attention of each pixel belonging to a class throughout the whole dataset, and use them to generate visualizations for each class and attribute. This visualizations allow for an spatial analysis for different attributes, for example on Figure 7.11, there is a clear distinction between some attributes attending the top parts and others the bottom parts of image, this makes sense considering the vegetation class includes both trees (upper half) and grass and bushes (bottom half) which have different incidence depending on the attribute. This result is consistent with the previous literature (Rossetti et al., 2019; F. Zhang et al., 2018).

Another interesting case to analyze is the road class, shown in Figure 7.12. As expected the average attentions show a triangular shape on the bottom of the images, which is how the road is shown in most images, but the boring and beautiful attributes put more attention on the top of the images, which only happens on images that are mostly road, such as tunnels and highways, combining this with the fact that road correlates negatively

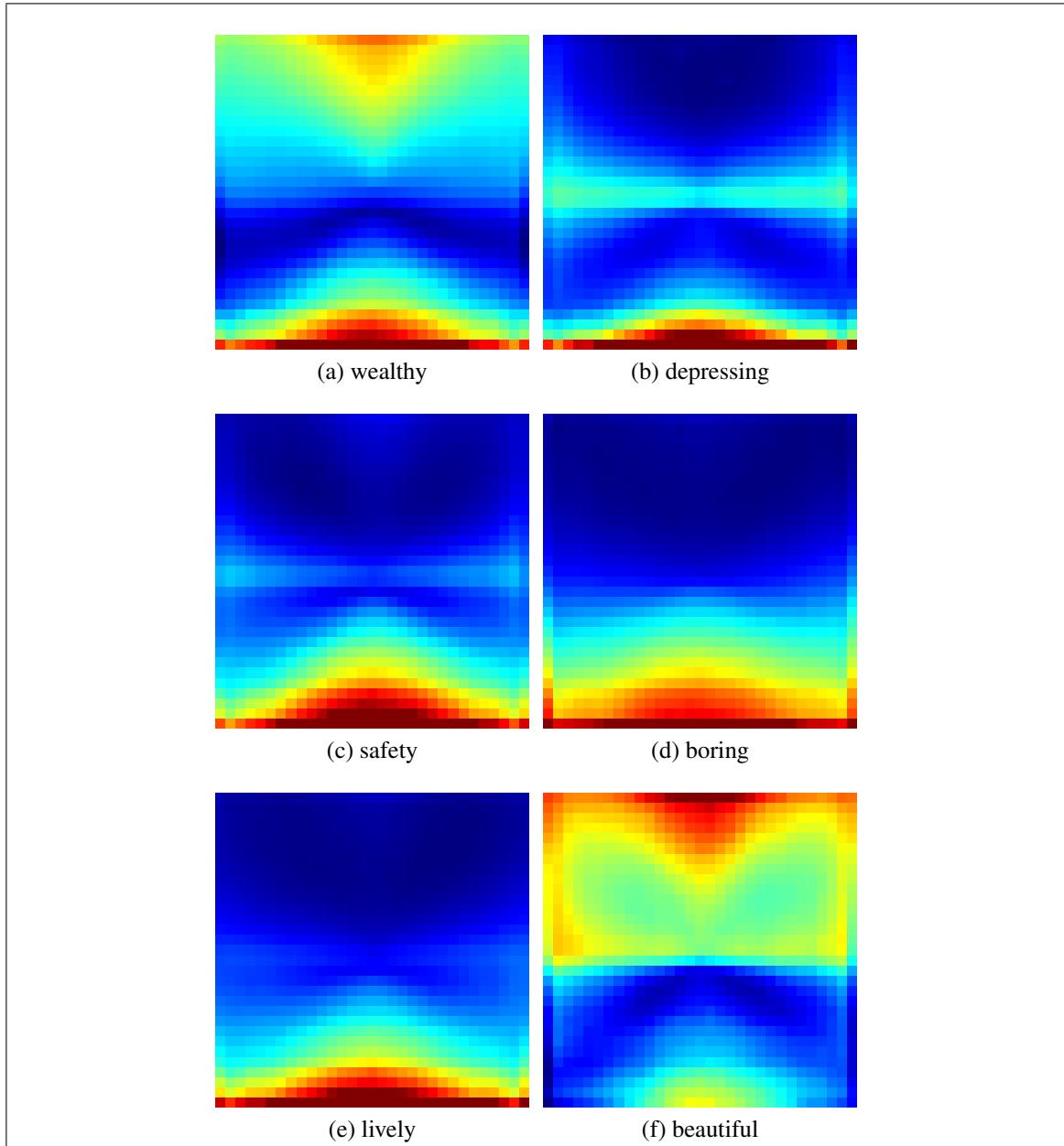


Figure 7.11. Attention averages for the vegetation class. Red means more attention. The color scales are logarithmic and are different for each image to ensure all the visualizations are clear.

with beautiful and very positively with boring, we can see that the results are reasonable since highways and tunnels are indeed not so beautiful and quite boring.

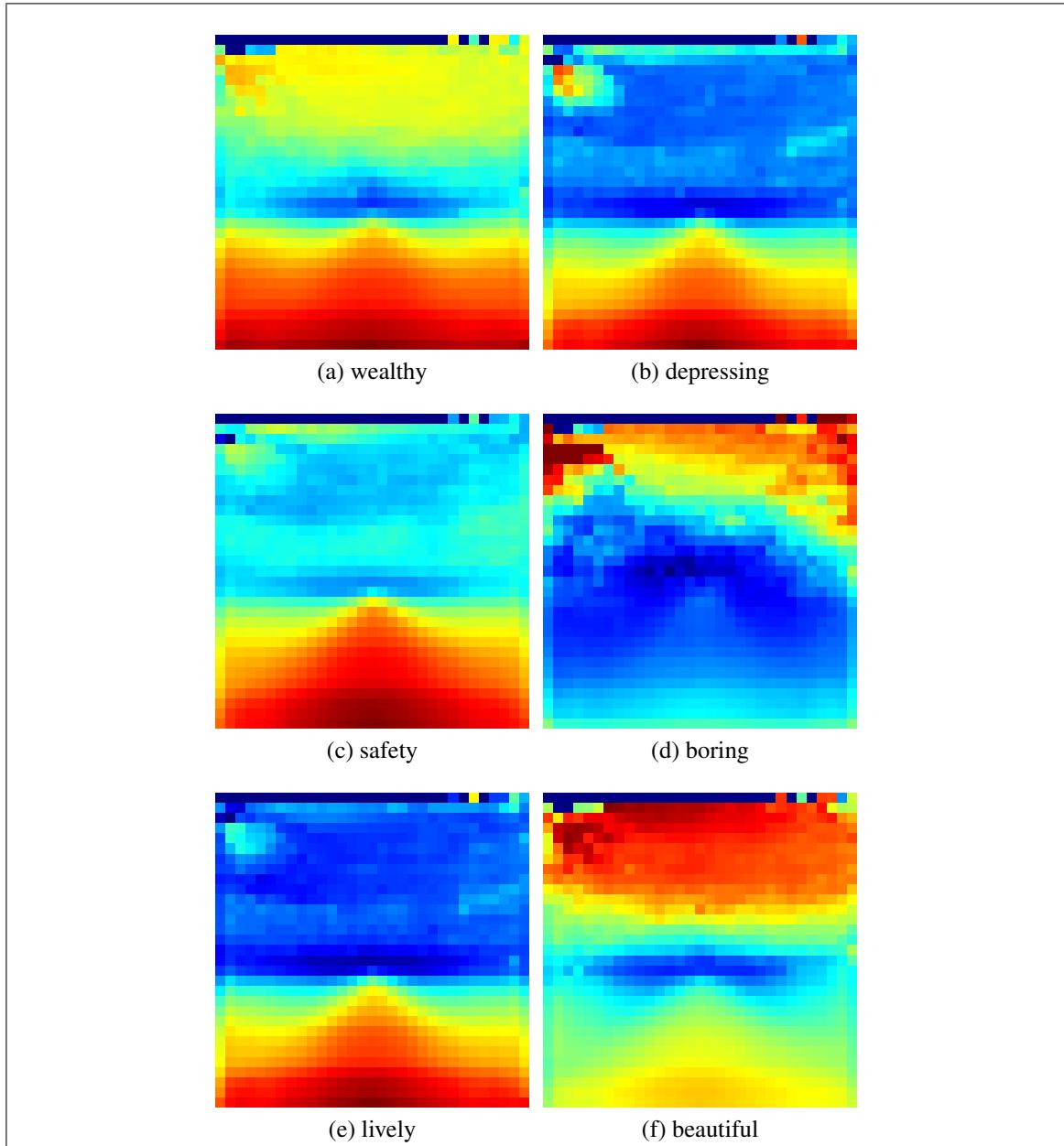


Figure 7.12. Attention averages for the road class. Red means more attention. The color scales are logarithmic and are different for each image to ensure all the visualizations are clear.

## 8. CASE STUDY

In order to do a practical evaluation of the model’s performance and an assessment of the model’s applicability, most of the recent literature include a case study for a particular city (Rossetti et al., 2019; F. Zhang et al., 2018, 2020; Quercia et al., 2014; Ordóñez & Berg, 2014; Liu et al., 2017). We continue this trend and use the AttnSegRank models trained on PlacePulse to analyze 120,000 Google Street View images of Santiago de Chile.

We use the results to generate a visualization of the city showing how the perception attributes behave throughout the different sectors. These are shown in figure 8.1, and it is easy to see how they precisely replicate the city’s actual income distribution, shown on Figure 8.2. Santiago is known for a considerably segregated urban distribution, in which the wealthier classes, and a large portion of goods and services are concentrated in a cone extending from the center of the city to the north east (Sabatini, Cáceres, & Cerdá, 2001), due to that the 6 attributes show a highly similar pattern that reflect that reality very well.

Some interesting insights are that as it was mentioned on section 7.5, highways and long roads are marked as very boring, which can be seen on the intense red lines present on the boring attribute, and that the more lively places (and the less boring) are much more concentrated towards the center of the city than the north east corner, which we believe is due to these sectors being the most busy in the city, with constant flow of cars and people at all times while also having a good amount of green areas.

In order to have a quantitative measure of the ranking generated for Santiago we calculate the mean score over the images of each commune in the city and compare the results with their respective socioeconomic indicators. We find a strong correlation between our perceived wealthy score and the poverty rate, and between our perceived depressiveness score and social vulnerability. We show these results on figures 8.3 and 8.4.



Figure 8.1. Urban Perception for Santiago de Chile, each dot represents an image that was analyzed with our model.

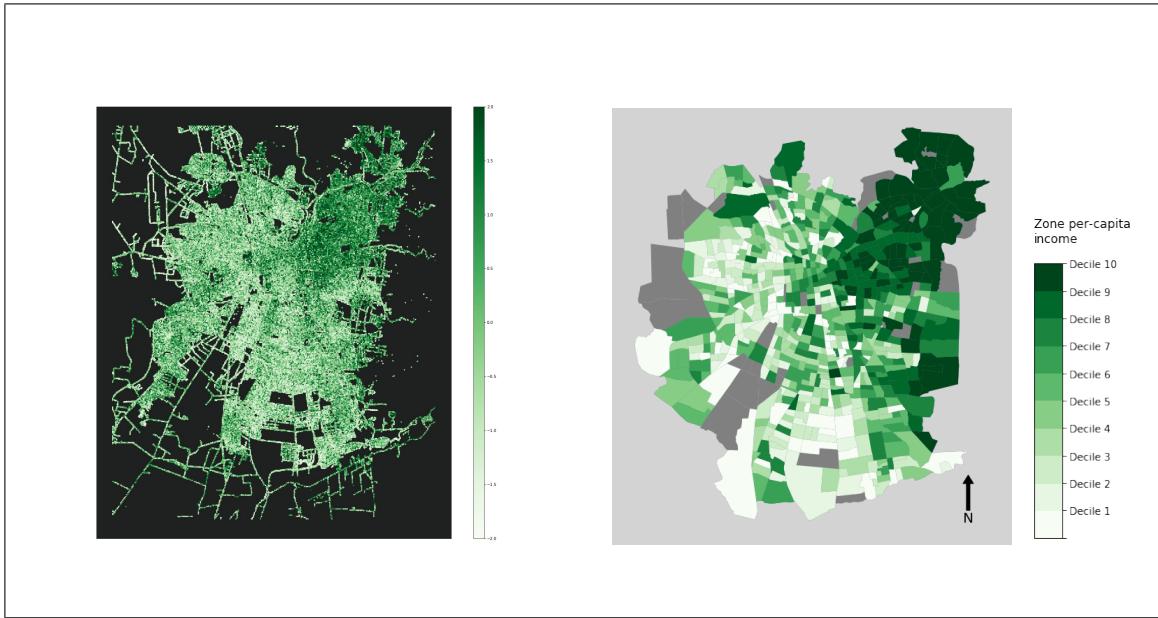


Figure 8.2. Wealthy score compared with the wealth distribution of Santiago de Chile by deciles. Darker means wealthier. Decile colormap reproduced with permission from Ramírez (2020).

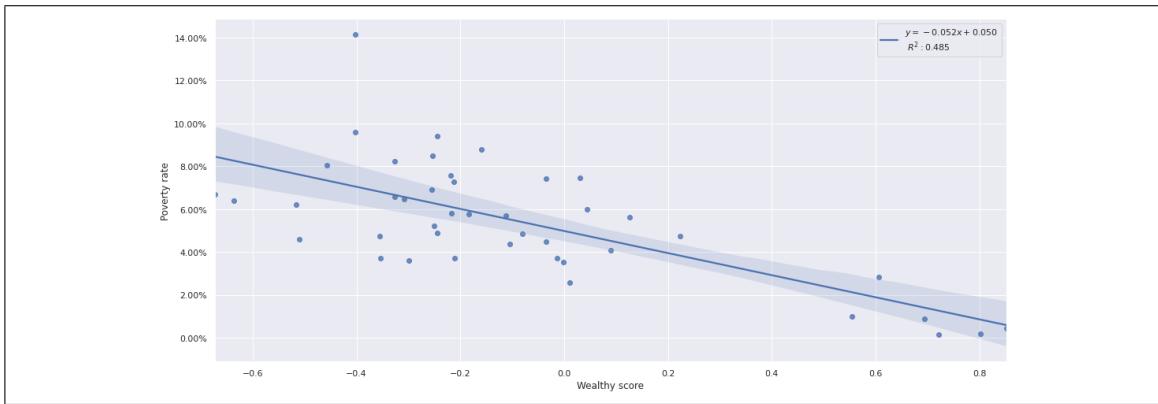


Figure 8.3. Poverty rate vs perceived wealthy score by commune in Santiago de Chile. Data taken from Ministerio de Desarrollo Social (2017).

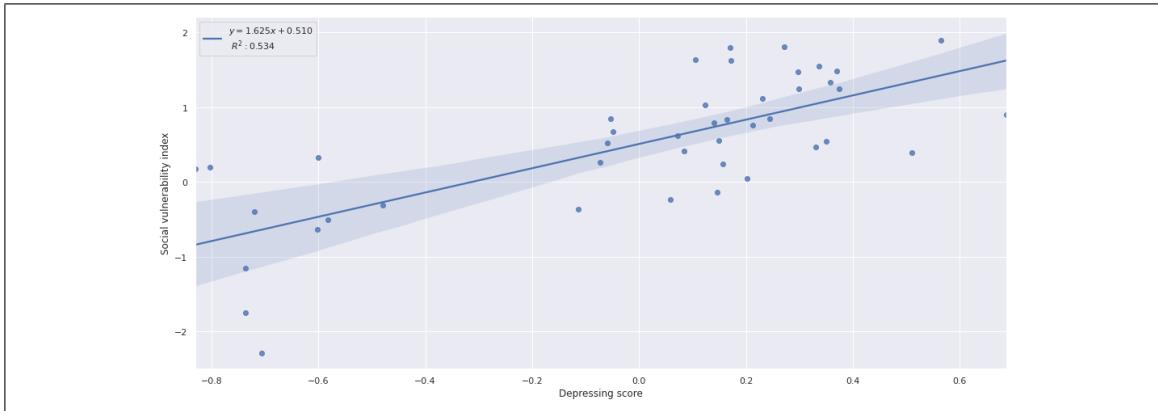


Figure 8.4. Vulnerability index vs perceived depressiveness score by commune in Santiago de Chile. Data taken from Asociación de Municipalidades de Chile (2016).

## **9. CONCLUSIONS**

Thanks to the massive increase on availability of large amounts of data and the advancements of deep learning, new ways to approach old problems have become possible in a wide range of fields. These models usually provide more effective and generalizable solutions, but take away a large portion of the result interpretability and the capacity to understand what the models are truly doing. The modelling of urban perception has made use of these advancements successfully, but the subjective nature of the problem makes it a task that is specially affected by the lack of explainability of modern machine learning algorithms. Due to that the recent literature has presented hybrid methods combining regression or discrete choice models with high level features extracted with pretrained neural networks such semantic segmentation or object detection. These techniques provide a better understanding of what is happening inside the model but sacrifice the higher expressiveness and performance that neural networks provide when trained end to end.

In this work we presented a novel neural network architecture, aimed at tackling this problem, through the use of semantic segmentation combined with standard deep learning methods like fine tuning convolutional features and attention mechanisms. This model is capable of successfully estimating the perception with a performance similar to those in the literature, but at the same time it outputs the attention weights and the segmentation of the image providing additional data that is human interpretable. We also present an aggregated analysis and visualizations of the results, that show that attention weights are a good tool for augmenting model explainability.

### **9.1. Contribution to the state of the art.**

#### **9.1.1. Enhancing explainability through high level features and attention.**

The main contribution of this research is an end to end trainable neural network architecture for urban perception quantification that presents very desirable explainability

properties. Unlike previous approaches in urban perception our models have the capacity to generate explainable insights on an instance level thanks to the semantic attention weights, making it a considerably more powerful tool both for research and practical application.

Additionally, the analysis on an aggregated level of segmentation and attention weights allows to draw similar conclusions to those from previous research, that based their models on more simple and interpretable techniques such as linear regression or econometric models, meaning that our architecture achieves an at least as good level of explainability but with a significantly more expressive and performing deep neural network.

### **9.1.2. Semantic segmentation as part of a neural network.**

The AttentionSegRank neural architecture that we propose, contributes with a novel way of combining semantic segmentation with traditional deep convolutional features through the attention operations, allowing the network to learn the capacity of dynamically choosing which parts of an image to attend based on the semantic classes of the segmentation. This idea is not exclusive to the urban perception task nor to ranking problems, and may be used for classification or regression in any computer vision task that allows for a pretrained segmentator to be used or that has available data to train a new one from scratch.

We think that this approach could prove useful on other vision domains where an improvement in model explainability is needed without sacrificing too much performance.

## **9.2. Future research directions.**

### **9.2.1. Improving dataset quality.**

Though it's very useful, the Place Pulse 2.0 dataset, it's very poor annotationwise and due to the way it was designed, changes on the google street view api are slowly

deprecating it. These problems make it hard for standardizing performance results and therefore making research that can be replicated easily. Generating a new dataset through a more strict process and with richer high level annotations, such as segmentation labels could open many doors for new research.

For example, experimenting if training the segmentator on the same dataset where the actual perception quantification will be learnt, or even training both things at the same time improves either performance or model explainability poses a very interesting research topic.

### **9.2.2. Additional high level features.**

There is also a possibility to experiment with other types of high level vision futures, such as object detection, which could serve either as a complement or as a replacement for semantic segmentation. Previous work has shown that object detectors are useful for quantifying urban perception (Rossetti et al., 2019), and therefore it remains an important open question if a deep learning model integrating an object detector in its pipeline could improve either performance or explainability.

### **9.2.3. Improved attention mechanisms.**

The attention mechanism used in this work is a simplified version of the now widely used transformer (Vaswani et al., 2017). Recent research has applied transformers and other similar attention mechanisms to several computer vision problems with great success. We abstained from using these techniques in our models because we made an effort to keep the models as simple as possible in order to reduce the black box nature of these very complex architectures. Improving over AttentionSegRank with a more modern, better performing attention mechanism and devising a way to keep a similar level of explainability with it, would be a very important contribution to the computer vision field.

## REFERENCES

- Adadi, A., & Berrada, M. (2018, 09). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access, PP*, 1-1. doi: 10.1109/ACCESS.2018.2870052
- Antonakos, C. L. (1995). Environmental and travel preferences of cyclists. *Transportation Research Part A: Policy and Practice*, 29(1), 85.
- Asociación de Municipalidades de Chile. (2016). *índice socio - delictual de nivel comunal en chile*.
- Badrinarayanan, V., Kendall, A., & Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR, abs/1511.00561*. Retrieved from <http://arxiv.org/abs/1511.00561>
- Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.
- Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the ieee international conference on computer vision* (pp. 3286–3295).
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. M., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *ArXiv, abs/2005.12872*.
- Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? an analysis of bert's attention. *CoRR, abs/1906.04341*. Retrieved from <http://arxiv.org/abs/1906.04341>
- Clifton, K., & Ewing, R. (2008, 03). Quantitative analysis of urban form: A multidisciplinary review. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, 1, 17-45. doi: 10.1080/17549170801903496

- Cordonnier, J.-B., Loukas, A., & Jaggi, M. (2019). On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., ... Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *CoRR, abs/1604.01685*. Retrieved from <http://arxiv.org/abs/1604.01685>
- Costa, G., Soares, C., & Marques, M. (2019). Finding common image semantics for urban perceived safety based on pairwise comparisons. In *2019 27th european signal processing conference (eusipco)* (pp. 1–5).
- Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). Imagenet: A large-scale hierarchical image database. , 248-255.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR, abs/1810.04805*. Retrieved from <http://arxiv.org/abs/1810.04805>
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (pp. 647–655).
- Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. *CoRR, abs/1703.04096*. Retrieved from <http://arxiv.org/abs/1703.04096>
- Doughty, H., Damen, D., & Mayol-Cuevas, W. (2018). Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 6057–6066).
- Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. *ArXiv, abs/1608.01769*.
- European union general data protection regulation (gdpr).* (2016). Retrieved from <https://gdpr-info.eu/>
- Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. In *Advances in neural information processing systems* (pp. 9273–9282).

- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *CoRR, abs/1512.03385*. Retrieved from <http://arxiv.org/abs/1512.03385>
- Huang, Z., Wei, Y., Wang, X., & Liu, W. (2019). *A pytorch semantic segmentation toolbox*. <https://github.com/speedinghzl/pytorch-segmentation-toolbox>.
- Iglesias, P., Greene, M., & Ortúzar, J. d. D. (2013). On the perception of safety in low income neighbourhoods: using digital images in a stated choice experiment.
- Jiang, B., Mak, C. N. S., Larsen, L., & Zhong, H. (2017). Minimizing the gender difference in perceived safety: Comparing the effects of urban back alley interventions. *Journal of Environmental Psychology, 51*, 117–131.
- Jiang, M., Chen, S., Yang, J., & Zhao, Q. (2020). Fantastic answers and where to find them: Immersive question-directed visual attention. In *Proceedings of the ieee/cvpr conference on computer vision and pattern recognition* (pp. 2980–2989).
- Johnston, A., & Carneiro, G. (2020). Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *Proceedings of the ieee/cvpr conference on computer vision and pattern recognition* (pp. 4756–4765).
- Khisty, C. J. (1994). Evaluation of pedestrian facilities: beyond the level-of-service concept. *Transportation Research Record, 1438*, 45-50.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*.
- Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Köppel, M., Segner, A., Wagener, M., Pensel, L., Karwath, A., & Kramer, S. (2019). Pairwise learning to rank by neural networks revisited: Reconstruction, theoretical analysis and practical performance. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 237–252).

- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.. Retrieved from <https://arxiv.org/abs/1602.07332>
- Laing, R., Davies, A.-M., Miller, D., Conniff, A., Scott, S., & Morrice, J. (2009). The application of visual environmental economics in the study of public preference and urban greenspace. *Environment and Planning B: Planning and Design*, 36(2), 355-375. doi: 10.1068/b33140
- LeCun, Y., Bengio, Y., & Hinton, G. (2015, 05). Deep learning. *Nature*, 521, 436-44. doi: 10.1038/nature14539
- LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541–551.
- Li, G., Wang, X., & Zhu, W. (2019). Perceptual visual reasoning with knowledge propagation. In *Proceedings of the 27th ACM international conference on multimedia* (p. 530–538). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3343031.3350922> doi: 10.1145/3343031.3350922
- Li, J., Tu, Z., Yang, B., Lyu, M. R., & Zhang, T. (2018). Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*.
- Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.
- Liu, L., Silva, E. A., Wu, C., & Wang, H. (2017). 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.
- Lynch, K. (1960). *The image of the city* (vol. 11). MIT press Cambridge, MA, USA.
- Marcel, S., & Rodriguez, Y. (2010). Torchvision the machine-vision package of torch. In *Proceedings of the 18th ACM international conference on multimedia* (p. 1485–1488). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/1873951.1874254> doi:

10.1145/1873951.1874254

- Ministerio de Desarrollo Social. (2017). *Resultados encuesta casen*.
- Nasar, J. L. (1990). The evaluative image of the city. *Journal of the American Planning Association*, 56(1), 41-53. doi: 10.1080/01944369008975742
- Noh, H., Hong, S., & Han, B. (2015, December). Learning deconvolution network for semantic segmentation. In *Proceedings of the ieee international conference on computer vision (iccv)*.
- Ordonez, V., & Berg, T. L. (2014). Learning high-level judgments of urban perception. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 494–510). Cham: Springer International Publishing.
- Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. Retrieved from <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- Porzi, L., Rota Bulò, S., Lepri, B., & Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd acm international conference on multimedia* (pp. 139–148).
- Quercia, D., O'Hare, N. K., & Cramer, H. (2014). Aesthetic capital: what makes london look beautiful, quiet, and happy? , 945–955.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 9.
- Ramírez, T. (2020). *Modelación de heterogeneidad en la percepción cualitativa de espacios públicos con modelos de elección discreta y aprendizaje de máquina* (Unpublished master's thesis). Pontificia Universidad Católica de Chile.

- Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Springer.
- Rose, J. M., & Bliemer, M. C. J. (2009). Constructing efficient stated choice experimental designs. *Transport Reviews*, 29(5), 587-617. Retrieved from <https://doi.org/10.1080/01441640902827623> doi: 10.1080/01441640902827623
- 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.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533–536.
- Sabatini, F., Cáceres, G., & Cerda, J. (2001). Segregación residencial en las principales ciudades chilenas: Tendencias de las tres últimas décadas y posibles cursos de acción. *eure (Santiago)*, 27(82), 21–42.
- Salesse, P., Schechtner, K., & Hidalgo, C. A. (2013, 07). The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*, 8(7), 1-12. Retrieved from <https://doi.org/10.1371/journal.pone.0068400> doi: 10.1371/journal.pone.0068400
- Santani, D., Ruiz-Correa, S., & Gatica-Perez, D. (2018). Looking south: Learning urban perception in developing cities. *ACM Transactions on Social Computing*.
- Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, 4(7), 170170.
- Shi, T., Zhang, X., Wang, P., & Reddy, C. K. (2020). A concept-based abstraction-aggregation deep neural network for interpretable document classification. *arXiv*, arXiv–2004.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56), 1929-1958. Retrieved from <http://jmlr.org/papers/v15/srivastava14a.html>

- Tompson, J., Goroshin, R., Jain, A., LeCun, Y., & Bregler, C. (2015). Efficient object localization using convolutional networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 648–656).
- Torres, I., Greene, M., & Ortúzar, J. d. D. (2013, 07). Valuation of housing and neighbourhood attributes for city centre location: A case study in santiago. *Habitat International*, 39, 62–74. doi: 10.1016/j.habitatint.2012.10.007
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR, abs/1706.03762*. Retrieved from <http://arxiv.org/abs/1706.03762>
- Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.
- 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.
- Zhang, F., Zu, J., Hu, M., Zhu, D., Kang, Y., Gao, S., ... Huang, Z. (2020). Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems*, 81, 101478.
- Zhang, Y., Niebles, J. C., & Soto, A. (2019). Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)* (pp. 349–357).
- Zhang, Z., Lan, C., Zeng, W., Jin, X., & Chen, Z. (2020). Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 3186–3195).
- Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid scene parsing network. *CoRR, abs/1612.01105*. Retrieved from <http://arxiv.org/abs/1612.01105>
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487–495).

Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., & Kankanhalli, M. (2019). Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th ACM international conference on multimedia* (p. 521–529). New York, NY, USA: Association for Computing Machinery. Retrieved from <https://doi.org/10.1145/3343031.3351040> doi: 10.1145/3343031.3351040

## **APPENDIX**

## A. SEMANTIC SEGMENTATION

### A.1. Visual representation

For visually representing segmentation, we make a color map over the images, following the cityscapes color palette (Cordts et al., 2016). See Figure A.1 for the exact palette and class list, and Figure A.2 for an example.



Figure A.1. Segmentation color palette.



Figure A.2. CityScapes sample.

## A.2. Segmentation distribution

Given the domain shift from Cityscapes to PlacePulse, is important to check how the segmentation behaves on the new dataset.

The most significant difference between the datasets is the image size, which are almost 13 times bigger in cityscapes, allowing for smaller objects like traffic signs to be clearly distinguishable. In training CS images are used with size of  $769 \times 769$ , while place pulse images are used with the standard  $224 \times 224$ . Another important difference is the origin of the images, while Cityscapes images were all taken on different cities of the same developed country (Germany), PlacePulse images come from 56 cities distributed on all continents, including both developed and developing countries, with the later ones contributing images with a significant visual difference.

Table A.1 show the percentage of pixels belonging to each segmentation class on the entire set of images of the PlacePulse dataset. Evidently this are not ground truth labels, but the ones obtained by our PSPNet trained on Cityscapes. As was expected, classes representing physically smaller objects have an almost negligible contribution since the smaller image size renders them pretty much unidentifiable. Domain shift makes the model constantly confuse the sidewalk (which can be seen in a large percentage of PlacePulse images) with the main road, reducing the class presence to a very underwhelming 0.96%. Same behavior can be perceived with the terrain class.

Segmentation Class	% of pixels
Building	26.60%
Vegetation	25.52%
Road	24.21%
Sky	6.24%
Fence	5.09%
Truck	2.96%
Car	1.94%
Person	1.48%
Bicycle	1.36%
Motorcycle	1.28%
Sidewalk	0.96%
Wall	0.89%
Terrain	0.65%
Pole	0.33%
Train	0.26%
Bus	0.14%
Traffic sign	0.05%
Traffic light	0.02%
Rider	0.02%

Table A.1. Pixel distribution of the segmentation classes over the PlacePulse Dataset

## B. SIGNIFICANCE TABLES

Here we present the full significance tables from the results discussed in Section 7.2

Class	Significance %						
	Wealthy	Depressing	Safety	Boring	Lively	Beautiful	Average
Road	62.46	<b>77.98</b>	<b>78.64</b>	31.77	<b>90.58</b>	17.33	<b>59.79</b>
Building	<b>63.74</b>	9.21	30.35	<b>82.69</b>	17.99	<b>52.79</b>	42.80
Sky	43.81	44.51	39.34	18.34	45.68	45.80	39.58
Person	4.86	18.18	10.83	28.83	25.76	6.77	15.87
Car	4.09	19.70	12.02	34.63	6.12	15.50	15.34
Vegetation	4.27	22.56	20.57	5.50	5.45	24.08	13.74
Bicycle	4.81	8.25	10.32	15.48	16.57	10.10	10.92
Sidewalk	10.65	12.46	2.21	7.65	7.35	9.66	8.33
Truck	4.65	11.38	2.94	9.57	6.29	14.29	8.19
Motorcycle	4.92	5.31	6.68	8.10	9.77	10.97	7.63
Wall	3.43	4.63	4.90	9.95	5.33	9.14	6.23
Fence	5.33	13.15	6.83	1.75	2.17	4.55	5.63
Pole	1.48	2.76	2.04	8.00	13.03	3.49	5.13
Terrain	0.74	2.72	2.72	13.31	1.45	4.38	4.22
Traffic sign	1.18	2.39	1.47	3.55	2.59	1.44	2.10
Traffic light	1.60	0.89	0.47	2.00	1.19	1.42	1.26
Train	0.22	1.57	0.16	0.73	0.52	3.84	1.17
Bus	0.74	1.51	1.15	0.82	0.83	0.92	1.00
Rider	0.49	1.12	1.02	1.10	1.23	0.98	0.99

Table B.1. Percentage of images where a segmentation class is considered significant over the entire PlacePulse dataset.

Class	Significance %						
	Wealthy	Depressing	Safety	Boring	Lively	Beautiful	Average
Sky	<b>78.49</b>	<b>80.12</b>	70.85	33.32	82.16	<b>82.62</b>	<b>71.26</b>
Road	64.87	81.01	<b>81.67</b>	33.06	<b>94.07</b>	18.00	62.12
Rider	30.34	65.75	59.74	61.42	71.21	59.14	57.93
Building	69.73	10.07	33.23	<b>90.59</b>	19.67	57.81	46.85
Wall	25.11	34.35	36.34	72.00	39.70	67.03	45.75
Motorcycle	23.73	25.34	32.13	38.80	46.53	51.87	36.40
Bus	26.46	53.96	40.25	27.17	29.32	31.90	34.84
Bicycle	14.42	24.77	30.73	45.93	49.26	30.07	32.53
Person	9.61	35.91	21.37	56.87	50.77	13.35	31.31
Car	6.53	31.41	19.16	55.41	9.73	24.66	24.48
Truck	13.13	31.95	8.26	26.92	17.51	40.09	22.98
Sidewalk	28.19	33.39	5.87	20.48	19.56	25.78	22.21
Traffic sign	12.00	24.43	15.18	36.81	26.60	15.08	21.68
Traffic light	26.92	15.04	8.06	34.47	20.28	24.78	21.59
Train	3.96	29.40	2.95	13.57	9.38	69.82	21.52
Terrain	3.35	12.39	12.40	61.04	6.63	20.23	19.34
Vegetation	4.87	25.70	23.45	6.26	6.21	27.42	15.65
Pole	3.89	7.19	5.35	21.04	34.17	9.15	13.46
Fence	10.88	26.98	14.06	3.61	4.47	9.35	11.56

Table B.2. Percentage of times a segmentation class is considered significant when it is present in an image.

## C. CORRELATION TABLES

This section shows the results from Figure 7.3 in a more detailed table format.

	Score and segmentation % correlation					
	Wealthy	Depressing	Safety	Boring	Lively	Beautiful
Bicycle	0.086866	-0.058904	-0.164551	-0.147775	0.160549	0.041282
Building	-0.280843	0.464943	-0.202526	0.134353	-0.171972	-0.508075
Bus	-0.120765	0.089133	-0.048565	-0.014197	-0.075356	-0.127071
Car	0.041143	0.045106	0.02467	-0.022377	0.070838	-0.041766
Fence	-0.24331	0.220695	-0.260908	0.061112	-0.172308	-0.218291
Motorcycle	0.02648	-0.000667	0.013527	-0.015124	0.036027	0.023063
Person	0.04549	-0.101353	-0.30913	-0.088583	0.058128	0.107055
Pole	-0.082125	0.053397	0.057317	-0.021622	-0.038195	-0.071639
Rider	-0.10875	0.071986	-0.048731	0.178123	-0.132248	-0.054794
Road	-0.000035	0.105378	-0.103487	0.216504	-0.10017	-0.122541
Sidewalk	-0.093546	0.050718	-0.077474	-0.009292	-0.078819	-0.048206
Sky	-0.090752	0.05207	-0.002863	0.146129	-0.154971	-0.015913
Terrain	-0.072509	0.013733	-0.101671	0.084799	-0.101522	0.017748
Traffic light	-0.028191	0.085843	-0.374202	0.057435	-0.038803	-0.035108
Traffic sign	-0.086816	0.050688	0.020622	0.017809	-0.075156	-0.066739
Train	-0.059097	0.112574	-0.073155	0.015524	-0.055187	-0.082677
Truck	-0.122791	0.193557	-0.124817	0.030997	-0.02044	-0.20529
Vegetation	0.368576	-0.616957	0.030223	-0.221661	0.212106	0.670974
Wall	-0.261629	0.216619	0.37146	0.150803	-0.230485	-0.190255

Table C.1. Correlation between segmentation percentage and correlation.  
Results for the AttnSegRank model.

	Score and attention % correlation					
	Wealthy	Depressing	Safety	Boring	Lively	Beautiful
Bicycle	0.057271	0.014189	-0.217687	-0.153383	0.126645	-0.021713
Building	-0.214182	0.484468	-0.22648	0.104895	-0.254365	-0.419682
Bus	-0.115311	0.11805	-0.085147	-0.027184	-0.086078	-0.16813
Car	0.003524	0.144843	-0.012164	-0.110474	0.039246	-0.088189
Fence	-0.233656	0.244746	-0.216828	0.076153	-0.203493	-0.258378
Motorcycle	-0.024055	0.052373	-0.012941	-0.029005	0.003378	-0.048763
Person	0.017417	-0.036531	-0.355335	-0.106455	0.046342	0.060095
Pole	-0.07281	0.135613	0.238478	-0.027945	-0.105173	-0.040957
Rider	-0.129007	0.200804	-0.044092	0.12438	-0.131043	-0.125526
Road	-0.07195	-0.164682	-0.104088	0.1565	0.176858	-0.191148
Sidewalk	-0.135462	-0.009194	-0.10564	-0.02815	-0.074386	-0.122792
Sky	0.054282	-0.053934	-0.05744	0.159083	-0.032063	0.183817
Terrain	-0.09714	0.060149	-0.084568	-0.041552	-0.094959	-0.054524
Traffic light	0.070855	0.154008	-0.382812	-0.009036	-0.116476	-0.035303
Traffic sign	-0.102794	0.15013	0.15488	0.003282	-0.1059	-0.096556
Train	-0.068101	0.101053	-0.122174	0.030762	-0.052124	-0.052673
Truck	-0.135761	0.237386	-0.126596	0.049528	-0.056697	-0.221666
Vegetation	0.336176	-0.575381	-0.038788	-0.164346	0.110463	0.648495
Wall	-0.226853	0.213753	0.295156	0.040262	-0.127957	-0.218601

Table C.2. Correlation between attention percentage and correlation. Results for the AttnSegRank model.