# FaceLift: An explainable urban transformation pipeline


## ABSTRACT

Deep learning has allowed remarkable insights when it comes to analysis of images. But the quality to understand what a model learns and does is still out or reach. In this paper we propose a Deep Learning driven processing pipeline called *FaceLift*, that allows us learn intangible concepts in urban settings like beauty, safety and richness. For the purpose of specificity, we conduct all our studies on the dimension of beauty in urban images, but the process is extensible to any annotated property. In addition to learning these concepts, the pipeline allows approximate transformation of street-view images so as to maximize or minimize these concepts in urban settings. We further design the pipeline using a novel explain-ability aimed architecture with a tandem of Generative Adversarial and Deep Neural networks. We validate our pipeline's Transformational capabilities using crowd sourced experiments and then evaluate the pipeline using several metrics, drawn from Urban design Literature. We conclude by summarizing actual expert insights about the use of such a tool and discuss the broader implications.



**ACM Reference format:**

. 2017. FaceLift: An explainable urban transformation pipeline. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 9 pages.

DOI: 10.1145/nnnnnnn.nnnnnnn


## 1 INTRODUCTION

Deep neural nets are progressing at an amazing pace over the past decade. The community and the technology has been breaking grounds across the fields. When it comes to classification and inference for specific tasks like object detection, scene detection, language modelling etc, Deep learning has been at the fore front. But the reasons behind a particular decision taken by a deep model, more or less still remains a mystery. This gap is by some accounts call the explainability gap, or the black box problem of Neural nets. Explainability and understanding the deep reasoning behind decisions is one of the most researched problems in the machine learning community.

The problem of explain-ability becomes even more abstract and obscured, when we are dealing with tasks that handle meta, and abstract quantities like sentiment, affects and aesthetics. Despite the black box like nature, deep neural networks have done remarkable strides in understanding creativity [24], memorability [12] or beauty [26] at a meta level, and works great in providing inferences. These works explore perceptual qualities of media objects using deep learning, and treat the explainability of the models using round about methods like perturbation of input and understanding correlation of several governing variables with decisions of the network etc. These methods are perfectly valid and do give some interesting insights into the decision influencing factors for the models, however still fail to explain the knowledge that the network has learn't and which it uses to drive the decisions that it takes.

Despite these issues, computational aesthetics have reaped a lot of benefits from the advances in Deep neural networks. Works like [14] [20] [17] [5], underscore the utility of machine learning in computational aesthetics in urban settings. In this paper we propose a generalizable pipeline for analysing geo-referenced images (FaceLift) for affective and/or aesthetic metrics. For the purpose of specificity, we conduct all our studies on the dimension of beauty in urban images. The pipeline learns the concept urban affect and then for a given street-view image finds an approximate transformation so as to maximize or minimize the presence of said affect. The is done by generating synthetic images to maximize the affect and then finding a best match to the synthetic images from the street-view database under some compositional constraints. We show that the humans overwhelmingly agree with the pipeline's transformations using a crowd sourcing experiment and then show that the transformations of the pipeline can be explained using well known urban design metrics.

## 2 RELATED WORK

This paper follows up on previous research in the field of computational aesthetics, namely that branch of computer vision dedicated to the analysis of visual quality and appeal, and, more in general, of subjective properties of visual data. Early work in the field done by Datta [4] looked at the beauty aspect of images using hand-crafted visual features and datasets collected from photo-contest websites. The introduction of deep-learning in this field boosted the activity. Works such as [12] used it to understand memorability. Some other works like [14, 26, 30] , explored the dimensions of beauty, aesthetics and their linkages to popularity and engagement over the web. The work by Redi [24] looked at quantification of the notion of creativity in the short microvideos. These works look at properties which are abstract and very subjective. But still they all claim impressive performances in these aspects. But all of them have a gap in explaining why their models have a good performance and what features have the classifiers learnt to look for. These questions boil down to the concept of explainability of machine learning models.

Similar to this paper, few works have focused on developing comptational aesthetics tools and datasets specific to urban perception. Works like StreetScore [17] and [25], collected and analyzed responses to images of urbanscapes across different subjective dimensions including safety, depression, beauty. An extension of this work [7] used deep learning to train models capable to rank urban images according to these subjective dimensions. Extending quantification of subjective information to the realm of maps was also explored by works such as [1, 19, 21, 22]. These works took the subjective dimension and overlayed on real world maps to present




*Conference'17, Washington, DC, USA*





DOI: 10.1145/nnnnnnn.nnnnnnn

| **symbol** | **stands for** |
|---|---|
| $X$ | Georeferenced urban images data |
| $I_i$ | Georeferenced image where $I_i \in X$ |
| $Y$ | Annotations classes for $X$ in affective space ( We work with Beauty ) |
| $y_i$ | Annotation class in $Y$ |
| $\hat{I_j}$ | Template image |
| $I'$ | Target Image |
| $C$ | Image Classifier |
| | |
| **term** | **stands for** |
| *Template Image* $\hat{I_j}$ | A synthetic transformation of input image $I$ towards the class $y_j$ |
| *Target Image* $I'$ | The natural image which is most visually similar to the template image |
| *Data Clustering* | A process which groups images in $X$ according to visual similarity (e.g urban vs rural) |
| *Data Augmentation* | A process which looks for images taken in the surrounding areas of the georeferenced images in $X$ |
| *Classifier* | A deep-learning framework that is able to classify images into one of the classes in $Y$ |
| *Generator* (*GAN*) | A deep-learning based generative framework to produce images similar to the ones in $X$ |
| $DGN - AM$ | A framework that, given the GAN and the Classifier, transforms an input image into the template image. |

**Table 1: Notations and Terms.**

a new dimension in which one can explore their world. The limitation of most of these works is that the models developed for urban perception are not interpretable, i.e., they provide predictions regarding subjective qualities of urban images, without explaining the reasoning behind the predicted score. Although very accurate, these models might not be useful in context of urban planning, since ... **[@Daniele: help?]**

The work that is the most similar to ours in terms of motivation and setup [27, 31] aim at prediction of scenic-ness of a place using a crowdsourced rated data. **[@Sagar the work from ATI is missing? Please add it and check differences. Mainly that we beautify instead of predicting and that ]**

The literature discussed here shows that there are gaps in understanding of the models that do very well when it comes to perceptual properties. On the other hand, we also see that there has been some progress in visualizing and understanding the internal reasoning of neural networks **[REF]**. We exploit recent developments in this field to **[...]**

## 3 FACELIFT PIPELINE

We present here a pipeline for transforming natural geolocated images from one category to another. The current process concentrates for Beautification of urban images for the sake of this study, but the pipeline is generalizable for any annotated urban image dataset. The components work equally effective given that the principles of training a deep-learning model are followed.

For the sake of berevity, we summarise the notations used in Table 1 and the pipeline steps in Figure 1.

In general terms, the framework allows any arbitrary set of *geolocated* images $X = I_1, I_2 .... I_n$ annotated in classes $Y = y_1, y_2, ..., y_k$, to transform natural images between classes: the algorithm can transform an image $I_i$ belonging to class $y_i \in Y$ , to image $I_j$ from class $y_j \in Y$. Both $I_i$ and $I_j$ are natural, non-synthetic images. Despite having another *meaning* (i.e. category), $I_j$ maintains the structural characteristics of $I_i$ (e.g. point of view, layout). This allows to visually reason about the discriminative properties between classes $y_i, y_j \in Y$, and visually understand the salient characteristics that drive a classifier to distinguish between classes $y_i, y_j$. The questions pertaining to what the network learns semantically have been explored for popular use cases [13, 16] but still remain largely unexplored for intangible classes representing concepts like beauty, sentiment etc. In this paper, we apply this general scheme to the specific problem of predicting beauty and explaining the changes that influence the beautification process.

The transformation framework consists of three phases (see Fig. 1). The first phase classifies images from $X$ into the corresponding categories $Y$ with high accuracy , using a convolutional neural network $C$.

The second phase transforms am image from class $y_i$ to class $y_j$, using Generative Adversarial Networks[23]. The output of this phase is a synthetic image $\hat{I_j}$, which summarizes the basic traits of the destination class $y_j \in Y$. The last phase matches the synthetic image $\hat{I_j}$, with the closest natural image in $X$. This matched natural image is then used to reason about urban design metrics that the network learnt to associate with beauty. We do this by formulating 5 urban design metrics in terms of different measured attributes.

For the rest of this section, we would delve deeper into the specifics of the pipeline that produces beautification of urban images.

### 3.1 Phase 1: Classifying Beauty

We design here a classifier $C$ able to correctly assess the beauty category $y_i$ of an image in $X$ using a deep learning network, we need first make sure we have enough reliable data to train the classifier. We do this by augmenting the available geolocated image data.

*3.1.1 Data .* Our seed dataset comes from StreetScore, a research work on urban affects scores [17]. The dataset in total contains 11183 images across the world from Google StreetView. The images are compared in a pairwise manner for qualities such as beauty, depression, richness , safety etc. For the purpose of our paper, we use the annotations for beauty. To train our classifier $C$ to detect beauty in terms of categories $Y$, we need to transform pairwise votes into absolute scores, then discretize absolute scores into a finite set of categories $y_i$. We transform the pairwise votes into ordinal scores using the TrueSkill [11] algorithm. To ensure reliability of absolute judgements, we filter out images with less than 3 votes. To discretize the resulting scores, we heuristically partition the data into two classes with maximum separation. Figure 2 shows the distribution of Trueskill score estimates with the threshold scores at which we decide beauty or ugly class boundary.

*3.1.2 Data Augmentation similarity bound.* Since smaller data size implies that a machine learning model has a risk of overfitting, we augment the data with some additional real and transformed data. Here, we take advantage of the geo-located nature of the images in our dataset. We also take advantage of the fact that some urban places in proximity, look quite similar to each other. Despite over a hundred thousand images in the original data, after filtering we are left with 20,000 images only. This is non-ideal to train classifiers with substantial number of parameters such as convolutional neural networks. Hence we choose to augment the dataset by exploiting the

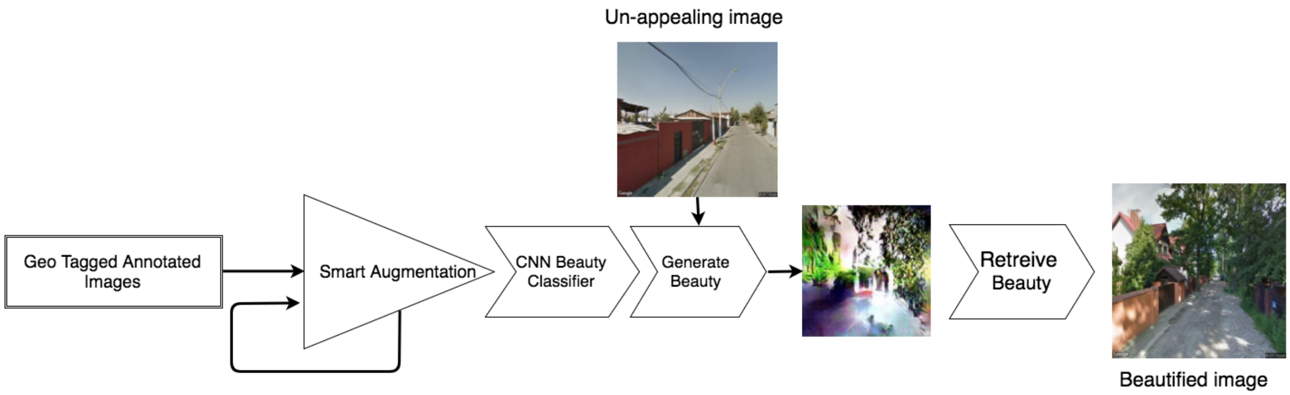



**Figure 1: Architecture of the Beautification Pipeline**

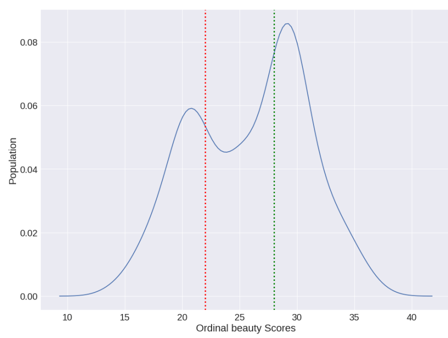



**Figure 2: Distribution of ordinal scores for images with at-least 4 votes. The red and green line represent the threshold below and above which images are tagged Ugly or beautiful. Images in between are dropped for separability**

geo-located nature of the image dataset. With an assumption that "**[A1]***rotation of camera, keeping the location constant, does not change the composition of the image considerably*" we now have two sets of images. First set is $R(i) \forall i \in I$ , which consists of images acquired by just rotating the pitch for a given annotated image. We rotate the camera across different values $\theta \in -30°, -15°, 15°, 30°$. For these images, the beauty score can be safely transferred from $i$. Next we acquire set of Images $T(i) \forall i \in I$, from a near geographic vicinity of a StreetView image $i$ in the StreetScore datase, by translating the streetview camera at distances of 10, 20 40 and 60 meters. We then represent each image from both sets, using features from the last but one layer of PlacesNet [32]. We then calculate a set of cosine similarities $S_t = \{s(i, T(i)) \forall i \in I\}$ between each original image $i$ and all images in the augmented set $T(i)$. We also calculate another set of cosine similaritie $S_r = \{s(i, R(i)) \forall i \in I\}$ Now with assumption **[A1]** , we only select translational images in the augmentation set where $s(i, T(i)) < median(S_r)$. This implies that the translational images $T(i)$ are just as similar to the original annotated image, as images acquired by rotation of camera. Hence we arrive at a similarity bound

$$\rho = median(S_r) \text{ where } S_r = \{s(i, R(i)) \forall i \in I\} \tag{1}$$

*3.1.3 Semantics of Augmentable images.* Given that we now had a similarity bound to decide whether to augment or not a particular image, we wondered whether certain types of scenes are more prone to augmentation compared to others. So we partitioned our data in two sets

- $setA$: contains images where the median similarity between translated images and the original image $s(i, T(i)) < \rho$ .
- $setB$: Images whose similarity with their translated set is farther apart i.e. $s(i, T(i)) > \rho$ .

We describe each image in both sets according to the scene depicted, by collecting the PlacesNet [32] labels with the top5 confidence scores. We then aggregate such labels at a set level by computing a TF-IDF metric. The resulting set of {label,Count} pairs are essentially how common or uncommon is a particular scene label in $setA$ compared to $setB$

The resulting prevalences of scene types can be seen in Fig 4. The plot shows that scenes like highways, fields and bridges, typically more uniform and open, don't change despite translation. On the other hand, there are more urban scenes like gardens, residential neighbourhoods , plazas and skyscrapers which are intuitively more sensitive to change in perspective by translation.

*3.1.4 The Beauty Classifier.* Once we have enough data, we train a deep convolutional network so as to classify images into the $Y$ beauty classes. One may use several successful deep convolutional neural network architectures, which work for other use cases like AlexNet [15] , PlacesNet [32] or GoogLeNet [28]. For our paper we use CaffeNet which is a modified version of AlexNet. This trained classifier is a important component in the next phase, which is generation of images.

We employ the above observations to train the beauty classifier model. We start with the base dataset of 20k images, as described in Section 3.1.1. We then progressively augment the data: first with rotation across 5 angles, then rotation with uniform translation for all images, and then rotation and translation for only images which satisfy similarity bound as evaluated as shown in Equation 1. We then train a Convolutional neural net model based on AlexNet architecture [28] on each of these augmentated datasets. The training is done on 70% split of the data and tested on the 30%.

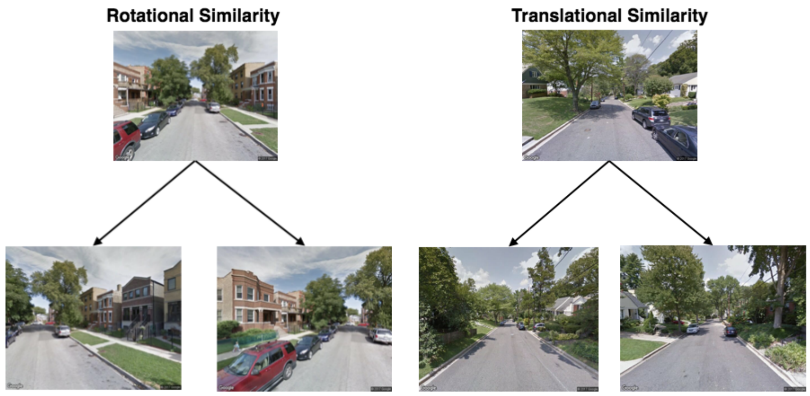



**Figure 3: Example images showing similarity of streetview scapes, when the camera is rotated by a small angle. The translational example shows similarity of images where the angle is less than the established bound $\rho$**

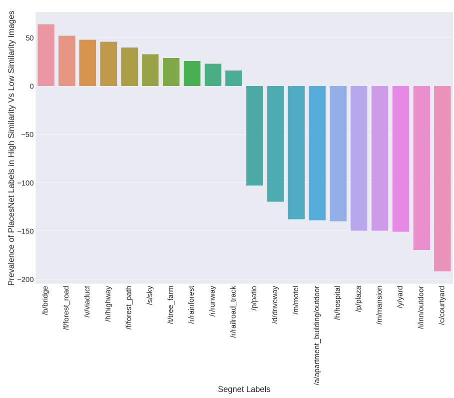



**Figure 4: Prevalence plot of types of scenes prevalent in Similar images compared to dissimilar ones.**

We see considerable improvement in classifier accuracy 2, with the best model performing at 73% accuracy for classifying images in two classes of Beauty and Ugly. This model represents the knowledge of the concept of beauty, learned from annotated and augmented streetview images.

| Policy | Accuracy (Percentage) |
|---|---|
| No augmentation | 63 |
| Rotation only | 68 |
| Rotation + translation | 64 |
| Rotation + Smart Translation | 73.5 |

**Table 2: Performance differences based on different augmentation policies, for the 3 Cities data**

## 3.2 Phase 2: Generating Images

We now want to design a framework to transform any image $I$ into a template image $\hat{I}_j$ (as shown in Figure 1) $\hat{I}_j$ is a synthetic version of the original image, with added features and motifs that maximize class $y_j$. To produce the template image $\hat{I}_j$, we need the following components in place,

- *Classifier*. We need a deep constitutional classifier that has learnt to classify between ugly and beautiful images. This implies that it has learnt the abstract concept of beauty to some extent
- *Generator*. We train a generative adversarial network (GAN) which can generate an approximate natural looking image drawn from distribution of a particular class of images, similar to the one in [6].
- *Activation Maximization*. We plug in the GAN and the classifier network into an Activation Maximisation (AM) framework. Given these components, an input image $I$, and a target beauty class $y_i$, the AM transforms $I$ in an ideal image $\hat{I}_j$ ( that maximizes the activation for beauty class $y_i$).

We have described the design and performace of Classifier in Sec. 3.1.4. We will delve deeper into the other two below

*3.2.1 Generator.* Generative Adverserial networks are an extremly useful tool when it comes to generating samples from a learned distribution[23]. GANs work by learning a pair of networks where one learns the distribution of sample space and generates samples using de-convolutional layers and another learns to discriminate between natural images from the training set and synthetic images generated by the generators. The problem is a min-max arrangement where we want to maximize error in discriminator by minimizing error between generated and natural images, there by generating samples that confuse the discriminator. But GANs are known to be very tricky to train [10] and hence to our end, we first try to use a pre-trained GANs on Imagenet images from [18]. We use this GAN to generate images for maximizing beauty, but because of the vast difference in the ImageNet image distributions for the 1000 classes, and the streetview images, the results were not very optimal. This provoked us to retrain the generator on the training dataset we used for Classification model. This improved the GAN performance considerably and it started generating images which do not entirely resemble natural streetview images, but look like paintings of these scenes.

#### 3.2.2 Activation Maximization.

We build on top of the activation maximization technique elaborated by Nguyen et. al [18] which utilizes the property of locality of codes with respected to generated images in Generator networks. Which means Generator codes which are close to each other would create similar looking images.This approach was initially aimed at visualizing the learnt knowledge of a convolutional neural network classifier. This is done by maximizing the activation of a particular output class probability neuron in a trained Classifier network, by feeding it images generated by a generator network. The maximization is achieved by doing gradient descent on the input generator codes with respect to the classifier neuron activation, keeping everything else locked. The result is a synthetic image that has a high activation for a pre-determined output neuron. To our end, we modify this method by starting the maximization method from a code which is closest to the a-priori input image, which is the ugly urban image $I_j$. This initializes the generator to a point from which the modified image should be ideally closest in terms of composition to the a-priori image. We then maximize the Beauty neuron of our trained classifier $C$ by doing the activation maximization process on the initialized generator, and stop as soon as the generated image pixels start getting saturated. The resulting output image is a natural-like image, which maximizes the beauty neuron for our classifier $\hat{I_j}$ . We hypothesize that because the process begins from an a-priori image, the resulting image is closest possible template to the ugly input image, but with the beauty neural activation maximized. Figure 5 shows the activation maximization output in the center.

## 3.3 Phase 3: Retrieving Images

In this final step we find a target image $I'$ from the dataset that is closely aligned, in terms of some visual similarity metric $E(I_1, I_2)$, with the generated template image $\hat{I_j}$ . The result of this exercise is to find the most similar looking image to an input image $I$ that maximizes a particular annotation class $y_j$. The visual differences in these two natural images, can act as the subject of reasoning for the explain-ability. The problem of finding images which are visually similar can be solved using image embeddings in a $N$ dimensional space $R^N$ We use a pre-trained deep network, which is trained to classify scene types to a very high accuracy [32] to extract the image embeddings. We use the last but one fully connected layer of this network to extract a 4096 dimensional feature vector from the template image. We then extract feature vectors from the complete test dataset using the same process. We can now use the $L_2$ Norm to find pairwise distances in the $R^{4096}$. Formally with $N$ test natural images and a template image $\hat{i}$ we extract $v_{\hat{i}} \in R^{4096}$ and and find pairwise distances $\{d_j \ \forall j \in N\}$ where $d_j = L_2(v_j, v_{\hat{i}})$ We then find the target image by finding the $min(\{d_j\})$. For the sake of redundancy, we find the top 5 such matches for every template $\hat{i}$ generated from every ugly image $i$. These target images are what we call the transformed images for maximizing beauty

## 3.4 Pipeline Validation

Because beauty is a subjective opinion, we need to understand if our pipeline is able to actually learn and generate the intangible qualities of beauty. For this, we run a user study to check how often humans agree with the machine's inference. For this reason, we take help of crowd-sourcing to understand how much do real humans agree with the pipeline's transformations. We randomly select 200 images, 100 beautiful and 100 ugly as per their TrueSkill scores. To have a reliable seperation in terms of visual appeal, we only select ugly images with scores less than 15 and beautiful images with scores greater than 30. This means that we are selecting images from the bottom 10 and top 10 percentiles according to the trueskill distribution as shown in figure 2. These images are then transformed to the opposite side of the spectrum of beauty using FaceLift. As a result, a beautiful image would be transformed into an ugly image and vice versa. Then we design an Amazon Mechanical Turk experiment, where we ask the turkers to choose the beautiful image between the original and the transformed images, without giving any hints of the transformation. We pay 0.1$ per human intensive task and we make sure that each Turker is a verified master, which assures that the Turker has a HIT approval rate above 90% for the past 30 days. We make sure that we have at least 3 votes on each image comparison, there by allowing us to choose majority voting. In all The results show that over all , the Turkers agree with the model **77.5%** of the time. Besides the over all agreement, the turkers agree **70%** of the time with the process of beautification and **85%** of the time with uglyfication. These results show that the facelift pipeline is learning the concept of beauty and then doing agreeable transformations on images.

# 4 VALIDATION METRICS

We found our pipeline to be an effective tool to beautify images of urban spaces. We now want to understand what the algorithm is looking at when transforming images. One way to do so would be to look at the template images and infer color and texture patterns. However, this approach is not scalable, as it would involve a substantial manual effort, and would be subject to personal interpretation.

One of the main contributions of this work is to develop metricsto explain what the network is learning as discriminatory features of beauty, in a fully automatic way. Table 3 shows a list of 5 evaluation metricstaken from urban design literature[2, 8]. These represent interpretable, measurable urban elements whose presence drives the aesthetic value of an urban environment. We design computer-based methods to map each of these theoretical metrics into a computationally measurable form. To measure the metrics, we select 500 ugly and 500 beautiful images from the test dataset based on their base TrueSkill scores. We then transform these images towards the opposite side of the beauty specturm using the FaceLift pipeline. We compare the values of these features between the two changed samples (beauty -> ugly , ugly->beauty), thus inferring statistics regarding which types of urban elements are added/removed by the beautification pipeline. In this Section, we present the computational methods used to map these 5 metrics, as well as the results that we get on the pre and post transform images.

## 4.1 Computational Methods

To validate the presence before/after transformation of the 5 urban elements in Table 3 we use a set computer-vision based tools as well as traditional data analysis techniques.

#### 4.1.1 Measuring Walkability and presence of Landmarks.

To quanitfy elements of Walkability, Greenery, and Landmarks, we

Original Template Retrieved

**Figure 5: Example of Beautification Process**

| Metric | Description |
|---|---|
| Walkability | Walkable streets are rated high on an aesthetic scale [8]. Walkable streets increase the social capital of a place and appeal to the exploring nature of human psyche. This implies that the urban space needs to address the fundamental need of people to walk and explore. This also implies that a walkable street must also be perceived as safe. |
| Green Spaces | Presence of Greenery is always pleasing to the eye. The literature always links urban beauty to curated and well maintained green spaces, where social interactions can happen [2]. This 'social' aspect of greenery implies that dense forests or unkempt greens are not always related to the sense of beauty in urban scenes. |
| Landmarks | Loosing a bearing in the city is not a very pleasant experience. Hence presence of recognisable and guiding landmarks influences the perception of an urban space [8]. |
| Privacy-Openness | A sense of privacy and a complimentary sense of openness are both influential in our perception of a place[8]. These values also tend to be related in an inverse 'U' fashion with beauty. |
| Visual Complexity | Visual complexity is a measure of how diverse a urban scene is. It manifests in terms of various design materials, textures and objects. Generally, visual complexity has an inverse 'U' relation with aesthetic values. The beauty and aesthetics of a place increases until it starts dropping because of 'too much' complexity[8]. |

**Table 3: Urban Design Metrics**

proceed as follows. We use PlacesNet [32] to extract information regarding the scene type (e.g. beach, garden, etc.). PlaceNet's output is the SoftMax distribution over 205 scene labels. We retain for an image the top 5 labels with higher confidence scores. We classify the 205 scene labels of PlacesNet into 4 categories, **L**andmarks , **A**rchitectural , **W**alkable , **N**atural. Each category is inspired from urban design literature [9]. Labels like *Abbey , Plaza , Courtyard, Garden, Picnic Area, Park , etc* fall into the category of *Walkable*, where as labels like *Mansion, Castle, Dam , Airport, etc* fall in the category of *Landmarks*. Labels like *Residential neighborhood, Motel, hotel, restaurant, etc* fall in the category of *Architectural* and labels like fields, pasture , forest, ocean, beach etc fall in the category *Natural*. For an image, we then measure its Wakability according to how many of the top-5 labels fall in category W. Similarly, we quantify presence of Greenery and Landmarks according to the frequency of N and L labels.

*4.1.2 Measuring Openness and Green spaces.* To measure Openness, we resort to Segnet [3], a semantic segmentation algorithm, which is trained on dashcam images from a real driving dataset , to detect 12 different elements in the image namely road, sky , trees , buildings , poles , signage , pedestrians, vehicles ,bicycles , pavement, fences and road markings. At the risk of over-simplifying, we can approximate that the openness of a street scene with the portion of sky visible in the scene, green cover by the portion of greenery detected in an image etc. We therefore quantify openness as the number of pixels labelled as 'sky' by segnet, and green cover as the portion of pixels labelled as greenery by segnet.

*4.1.3 Measuring Visual Complexity.* Visual complexity is a measure used in urban design measurement [8] to understand the diversity of a particular place. Ideally the complexity has a more granular nature, right from the texture of the roads and walls, to the groomed gardens or lack thereof. Again with a risk of over-simplication, but to approximate a computational metric, we define complexity as the amount of disorder in terms of distribution of urban elements in the scene. As described before, we use SegNet [3] , to extract a 12 dimensional stochastic vector consisting of the proportion of pixels belonging to each element for a given image. For a given image, we store these proportions into a stochastic vector $XH(X)$ on that vector:

$$H(X) = -\sum p(i) \log p(i) \tag{2}$$

The $i$ in Eq 2 is the SegNet dimension for one of the 12 objects. The entropy value $H(X)$ would become a proxy for visual complexity of the urban scenes. What we want to understand is not the absolute nature, but the trend in the variation of this value across the beautification process.

## 4.2 Testing Hypothesis

We now try to reason about different hypothesis inspired from the literature [2, 8, 9] using the computational approaches mentioned above. The main outcome of this excercise is either to confirm or refute whether we can explain how our model understands beauty using popular urban design concepts. Without the loss of generalization, we want to probe whether the elements that the pipeline deems beautiful or ugly, are grounded in literature.

*4.2.1 Walkability and Green spaces.* As mentioned in table 3 , Walk-ability of streets has high impact on the beauty and other aesthetic qualities of a place. We test this by quantifying the amount of greenery and walkable elements added by the beautification process. First, we transform 500 images from both sides of the beauty spectrum to either beautiful (if ugly) or ugly (if beautiful) urban scenes.

After detecting LAWN categories through PlacesNet labels, we compute, for each image, the difference between the category frequency before and after transformation (e.g. how many 'Walkable' labels are added after beautification?). We then plot the aggregated difference-distributions for beautified and uglified image sets in Fig 6.

*[H1] Walkable streets is favoured in beautiful urban scenes*

To test that *H1* is valid, we first plot a prevalence count of different categories of labels for Beautified and Uglified images. It can be seen from Fig 6, that walkable scene types are highly favoured in the beautification process. Ugly images are transformed into Walkable spaces almost twice as frequently in beautification compared uglification.

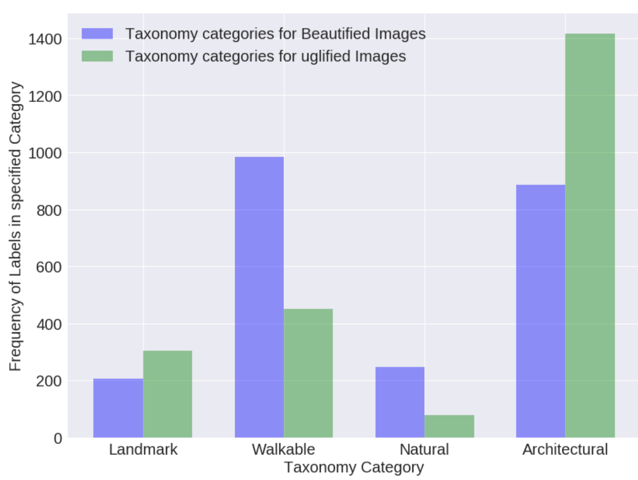



**Figure 6: Prevalence plot of categories of scenes prevalent in Beautified against Ugly-fied images**

*[H2] Green spaces are favourable for beauty in urban scenes.*

Figure 6 and 7 implies that natural scenes are twice as likely in beautified images than in uglified images. To test this hypothesis further, we further analyze the percentage of 'tree' pixels (according to SegNet) added by the beautification process: in average **[XXX% pixels]** of greenery are added after beautification.

*4.2.2 Privacy and Openness .* From the literature, it is conjectured that privacy is great when one looks at personal spaces, but when it comes to public settings, there is an inverse 'U' relation with how private a place feels like. Too much privacy discourages the

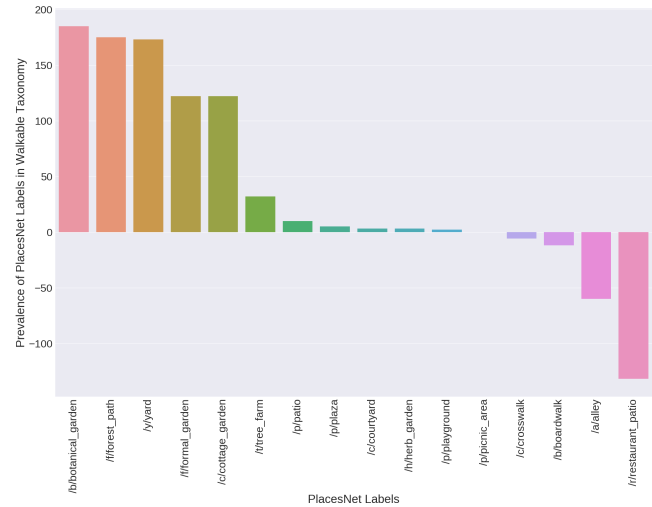



**Figure 7: Prevalence of Walkable labels in Beautified images against Ugly**

fundamental human urge to explore a mystery. Too much openness alerts the primal urge to feel safe.

*[H3] Sense of Privacy has an inverse 'U' relationship with the sense of beauty*

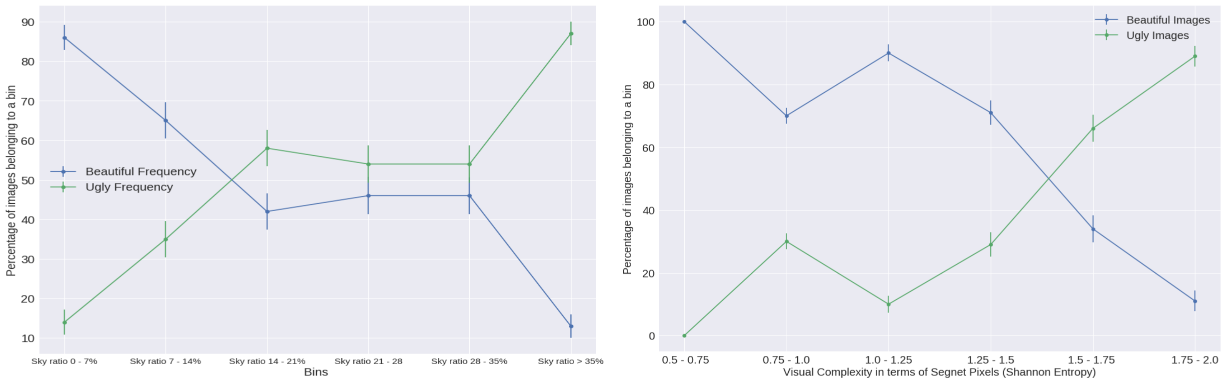
What *H3* suggests is that sense of privacy is not always associated with beauty. To understand the relationship between openness and beauty in urban scenes, we employ a technique called binned plots **CITE []**. We bin the range of sky pixels into **XXX Bins**Each image is then assigned to the bin corresponding to its proportion of sky pixels. Among the 1000 images (500 uglified, 500 beautified) in our data, we then repeatedly sample 100 images across bins, and count how many of the sampled images fall in either beautified or uglified transformed category. We plot the mean and standard deviation in a plot for these occurrence frequencies. The resulting plot gives a trend about how likely is the presence of pixels favoured or disfavoured by the beautification process.

It can be seen that our model prefers lack of openness in the beautification process. The inverse 'U' Relationship is completely absent and cozy urban places are actually favoured in the beautification process.

*4.2.3 Visual Complexity .* Visual complexity is a metric to measure the diversity of an urban scene. There is a trade off when it comes to balancing visual complexity with beauty. Too much diversity overwhelms the cognition and makes it hard to establish bearing.

*[H4] Visual Complexity has a inverse 'U' relation with the sense of beauty*

To test this hypothesis, we compute the binned plot (see Sec. 4.2.3) of our complexity metric. It can be seen from Fig 8b that visual complexity does peak in beautiful images but then deteriorates rapidly.




**(a)** **(b)**

**Figure 8: Figure 8a shows the Binned Plot for Sky pixels across transformed images and Figure 8b shows the Binned Plot for Visual Complexity across transformed images**

## 4.3 Interdependence of Urban elements

To understand the predictability and interdependence between the most influencing objects in an image and the probability of finding an image beautiful, we adapt the approach as described in [29], which proposes using logistic regression coefficients as a measure for upper bounds on influence of a variable. This method is also helpful in understanding the interdependence of variables for a particular outcome. Using the approach we perform a logistic regression over the two variables $V_1$ and $V_2$ which denote the ratio of a particular type of object pixels to the total area of image in pixels. So these ratios basically represent how much of the total image area is dominated by a particular object. The objects in this study are limited to the 12 urban object labels supported by SegNet [3]. Assuming dependence, we introduce a third term, which represents the factor that measures the dependence of $V_1$ and $V_2$ and is simply the product $V_1 * V_2$. The logistic regression would try to fit a line

$$L = invLogit(\alpha + \beta_1 * V_1 + \beta_2 * V_2 + \beta_3 * V_1.V_2) \quad (3)$$

Here $invLogit$ is the inverse Logistic function. According to the rule of fourth described in [29] , the coefficients $\beta_1, \beta_2 and \beta_3$ represent some properties about the influence of $V_1, V_2$. An increase in 1% area of the amount of pixels belonging to object $V_2$ would correspond to a maximum increase/decrease of $\frac{\beta_2}{4}$ percent in the likely hood of the image being beautiful. Same rule applies for $\beta_1$. In case of $\beta_3$ the co-efficient corresponds to mutual dependence. An intuitive explanation is that if a 1% change in value of $V_1$ would add $\beta_3$ to the value of $\beta_2$. Hence $\beta_3$ links changes in variable $V_1$ to changes in influence of variable $V_2$ and vice versa. We perform the logistic regression on the 5 prime influences namely *Sky* , *Buildings* , *Road* , *Vehicles* , *Trees*. The results of pairwise regression along with the dependency term are summaries in the table 4

| Object pair | $\beta_1$ | $\beta_2$ | $\beta_3$ | Error Rate (Percentage) |
|---|---|---|---|---|
| Roads - Vehicles | -0.015 | -0.05 | 0.023 | 40.6 |
| Sky - Buildings | -0.08 | -0.11 | 0.064 | 14.4 |
| Sky - Trees | 0.03 | 0.11 | -0.012 | 12.8 |
| Buildings - Trees | -0.032 | 0.084 | 0.005 | 12.7 |
| Roads - Trees | 0.04 | 0.10 | -0.031 | 13.5 |
| Roads - Buildings | -0.05 | -0.097 | 0.04 | 20.2 |

**Table 4: Regression coefficients**

# 5 DISCUSSION

## 5.1 Beyond Beauty

The pipeline is generalizable for geotagged and annotated images. The aim of this paper is to propose a pipeline with uses state of art methods in generative models to understand affects in urban images. But the pipeline can be easily extended towards a different outcome variable such as safety, public health, political status etc.

## 5.2 Limitations and biases

Like any supervised deep learning based framework, this work is only able to learn what is present in the data. Hence the method of acquiring annotations for urban images can introduce huge biases in the model. The current model is trained on images acquired from the study on streetscore [17]. However their annotation is open to general public and there is not way we can remove biases that come with culture and location, in a highly subjective effect like beauty. Moreover because the pair wise choice is simply done by clicking one of the two images, the data might have noise introduced by non-serious participants. Such biases are bound to be picked up by the deep learning model. One can argue that the preference of our model for greenery , is a form of bias in the data. Another Limitation of our work is in the metric formation. The computational metrics developed to capture the real urban design metrics are designed using heuristics. There needs to be more crowd and expert validation to establish the validity of their formulation.

## 5.3 Future work

# REFERENCES

[1] Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, and Francesco Aletta. 2016. Chatty maps: constructing sound maps of urban areas from social media data. *Royal Society open science* 3, 3 (2016), 150690.

[2] Christopher Alexander, Sara Ishikawa, Murray Silverstein, Joaquim Romaguera i Ramió, Max Jacobson, and Ingrid Fiksdahl-King. 1977. *A pattern language*. Gustavo Gili.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2015. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561* (2015).

[4] Ritendra Datta and others. 2008. Algorithmic inferencing of aesthetics and emotion in natural images: An exposition. In *2008 15th IEEE ICIP*. IEEE, 105–108.

[5] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. 2012. What makes paris look like paris? *ACM Transactions on Graphics* 31, 4 (2012).

[6] Alexey Dosovitskiy and Thomas Brox. 2016. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4829–4837.

[7] Abhimanyu Dubey, Nikhil Naik, Devi Parikh, Ramesh Raskar, and César A Hidalgo. 2016. Deep learning the city: Quantifying urban perception at a global scale. *arXiv preprint arXiv:1608.01769* (2016).

[8] Reid Ewing and Otto Clemente. 2013. *Measuring urban design: Metrics for livable places*. Island Press.

[9] Clemente Otto Ewing Reid. Measuring Urban Design - Metrics for Livable Places. (????).

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. 2017. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028* (2017).

[11] Ralf Herbrich, Tom Minka, and Thore Graepel. 2007. TrueSkillâĎć: a Bayesian skill rating system. In *Advances in neural information processing systems*. 569–576.

[12] Phillip Isola and others. 2011. What makes an image memorable?. In *IEEE CVPR*. 145–152.

[13] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[14] Aditya Khosla and others. 2014. What makes an image popular?. In *Proceedings of the 23rd WWW*. International World Wide Web Conferences Steering Committee, 867–876.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[16] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090* (2014).

[17] Nikhil Naik, Jade Philipoom, Ramesh Raskar, and César Hidalgo. 2014. Streetscore-predicting the perceived safety of one million streetscapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 779–785.

[18] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. 2016. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*. 3387–3395.

[19] Daniele Quercia. 2015. Chatty, Happy, and Smelly Maps. In *Proceedings of the 24th International Conference on World Wide Web*. ACM, 741–741.

[20] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. 2014. Aesthetic capital: what makes London look beautiful, quiet, and happy?. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 945–955.

[21] Daniele Quercia, Rossano Schifanella, and Luca Maria Aiello. 2014. The shortest path to happiness: Recommending beautiful, quiet, and happy routes in the city. In *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 116–125.

[22] Daniele Quercia, Rossano Schifanella, Luca Maria Aiello, and Kate McLean. 2015. Smelly maps: the digital life of urban smellscapes. *arXiv preprint arXiv:1505.06851* (2015).

[23] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434* (2015).

[24] Miriam Redi and Others. 2014. 6 seconds of sound and vision: Creativity in micro-videos. In *Proceedings of the IEEE CVPR*. 4272–4279.

[25] Philip Salesses, Katja Schechtner, and César A Hidalgo. 2013. The collaborative image of the city: mapping the inequality of urban perception. *PloS one* 8, 7 (2013), e68400.

[26] Rossano Schifanella and others. 2015. An Image is Worth More than a Thousand Favorites: Surfacing the Hidden Beauty of Flickr Pictures. In *Proceedings of THE 9TH ICWSM 2015*.

[27] Chanuki Illushka Seresinhe, Tobias Preis, and Helen Susannah Moat. 2017. Using deep learning to quantify the beauty of outdoor places. *Royal Society open science* 4, 7 (2017), 170170.

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[29] Brandon K Vaughn. 2008. Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J. *Journal of Educational Measurement* 45, 1 (2008), 94–97.

[30] Yilin Wang and others. 2015. Unsupervised Sentiment Analysis for Social Media Images. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI'15)*. AAAI Press. http://dl.acm.org/citation.cfm?id=2832415.2832579

[31] Scott Workman, Richard Souvenir, and Nathan Jacobs. 2016. Understanding and Mapping Natural Beauty. (2016), 5589–5598. DOI: http://dx.doi.org/10.1109/ICCV.2017.596 arXiv:1612.03142

[32] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.