EUROGRAPHICS 2017 / L. Barthe and B. Benes
(Guest Editors)

Volume 36 (2017), Number 2

On Realism of Architectural Procedural Models

J. Beneš[1,2], T. Kelly[2,3], F. Děchtěrenko[4], J. Křivánek[1], and P. Müller[2]

[1]Charles University, Prague
[2]Esri R&D Center Zurich
[3]University College London
[4]Institute of Psychology, Academy of Sciences of the Czech Republic

Abstract

The goal of procedural modeling is to generate realistic content. The realism of this content is typically assessed by qualitatively evaluating a small number of results, or, less frequently, by conducting a user study. However, there is a lack of systematic treatment and understanding of what is considered realistic, both in procedural modeling and for images in general. We conduct a user study that primarily investigates the realism of procedurally generated buildings. Specifically, we investigate the role of fine and coarse details, and investigate which other factors contribute to the perception of realism. We find that realism is carried on different scales, and identify other factors that contribute to the realism of procedural and non-procedural buildings.

Categories and Subject Descriptors (according to ACM CCS): I.3.5 [Computer Graphics]: Computational Geometry and Object Modeling—Geometric algorithms, languages, and systems

1. Introduction

Procedural modeling is a set of methods for generating computer graphics content. These methods ordinarily strive to achieve a key factor – realism – to provide a more engaging experience for the user and to increase the value of the final product.

Procedural methods typically approach the assessment of realism by qualitatively evaluating a small number of results. Occasionally a user study that compares the method's results and real world exemplars is employed. However, in general, there is a lack of systematic treatment and understanding of what is considered realistic in procedural modeling. Similarly, only a few studies investigate the realism of content in computer graphics.

We contribute to the discussion of realism by conducting a user study that investigates the realism of *procedurally* generated buildings. Specifically, our study set out to determine 1) the contribution of both fine and coarse details to the perception of realism and 2) the qualitative factors that contribute to the perception of realism.

We have found that the perception of realism of procedural buildings depends on different scales of detail in statistically significant ways, detailed in Section 5.1. In addition, we have identified a number of other factors, listed in Section 5.2, that contribute to the realism of procedural buildings. We discuss how these results generalize to non-procedural buildings in Section 6.

Our contributions are 1) proving the importance of coarser scale structure on realism, 2) identifying features that increase or diminish the realism of procedurally and non-procedurally generated buildings, and 3) a novel methodology for investigating procedural models using cut-outs to compare generated objects with reference photographs.

2. Previous Work

To the best of our knowledge, there is no systematic treatment of realism for *procedural rules or models*. Many works on procedural modeling explicitly attempt to achieve realism, e.g. [DHL*98, Ebe03], and many more mention it.

An editorial by Reinhard [REKS13] remarks that even low-resolution, "tourist" photographs preserve plausibility and proposes that it is the lack of fine detail that hinders the realism of synthetic (not necessarily procedural) images. One approach to remedy this lack of detail is CG2Real [JDA*11], an image-based rendering method that replaces parts of generated images with parts of photographs. The work also presents a small study to evaluate the realism of its results that is similar to ours. An interesting study into the nature of realism, albeit for photographs in a rendering and setup-of-scene context, was presented by Rademacher [RLCW01], who found that shadow softness and surface roughness have a statistically significant contribution to the perception of realism.

The procedural models that we investigate are typically created by generative procedural modeling approaches, such as shape-grammars [S*80], split-grammars [WWSR03, MWH*06, SM15], data-driven techniques that synthesize buildings from real-world data [FW16], or from a set of pre-defined parts [KCKK12, TYK*12]. However, non-automatic, user-guided techniques, e.g. sketch-based building generation [NGDA*16], also make use of information on the realism of procedural buildings.

Closely related, although not directly applicable, are image processing and machine learning approaches [DSG*12, MMW*11] that could help identify features missing from, or misrepresented on, a procedural model. Furthermore, there is a body of work concerned with image equivalence. Image Quality Metrics [Čad08] assess the arithmetic equivalence of images, whereas the Visible Differences Predictor [Dal92] is the primary method for perceptual equivalence. In contrast, we look for similarity in visual appearance, much like [RFWB07, RBF08]. Finally, the Structural Similarity (SSIM) metric [WBSS04] is used to assess the degradation of an image with respect to a reference image.

3. Design Considerations

Our experiment compares images of *procedurally* generated buildings to photographs. We aim to better understand the factors which contribute to the perceived realism of procedurally generated buildings, and buildings in general. We are not interest in photo-realism in an image equivalence sense, e.g., Visible Differences Predictor [Dal92], but rather in the quality of the generated models and their materials.

Below, we explain the procedural aspect of our experiment, discuss whether to run the experiment with or without reference, whether to run it in 2D or 3D, and the effects our experiment investigates.

Procedural Aspect. In procedural modeling, authors create procedural rules [MWH*06, SM15] that are expected to generate a wide range of building style variations. This makes procedural buildings susceptible to errors that are different from errors one would expect from manually modelled buildings; for example, incorrectly repeated textures, misplaced geometries, or repetition between models. Additionally, a procedural rule's output relies on the quality and variety of assets – one cannot expect to generate a realistic procedural city with only a limited number of textures – and authors may underestimate the difficulty of attaining realism across a rule. Other factors, such as how consumers would perceive procedurally generated buildings over prolonged periods, or when seen in aggregates [RBF08], play a role.

The importance of these factors is an insight that studying solely the realism of manually generated buildings could not provide. Similarly, performing this study with manually generated buildings may result in different factors being detected. While our design does not aim to investigate all of the above mentioned factors, it is nonetheless designed for procedural rules and conducted on procedurally generated buildings. Consequently, its results are more readily applicable to procedural rules. Applicability of the results to non-procedural buildings in discussed in Section 6

No Reference vs. With Reference. We chose to run our experiment with reference. While simply asking how realistic a single image is is tempting, it can also be misleading, as no base-line is given and because participants might operate with different notions of realism [RLCW01]. Other formulations, such as asking how photographic or synthetic an image appears, or explaining what to look for, could introduce preconceptions and bias the experiment.

2D Images vs. 3D Models. Instead of comparing procedural rules and their generated models in 3D, we compare their *images* (photographs and computer generated renders). The main benefit of this approach is that it is considerably easier to acquire and display the reference – in our case, photographs.

The main benefit of choosing to compare 3D models would be that the participants could view the objects from arbitrary angles, but there would be several drawbacks. The acquisition of a *real world* reference would be significantly more difficult (3D scanning) and perhaps imperfect. It is also not clear how occluded parts of objects would be acquired and displayed. The 3D reference model could be presented to the participant as a diffusely lit, solid color object. Though, in discarding texture and material information we would also lose a large part of what makes a 3D model realistic. Alternatively, it could be presented with its acquired texture, displayed using the same rendering method we would use to display the procedural model. However, this would limit us to real-time rendering methods with a stronger bias, or force us to revert back to 2D.

For the above reasons, we consider 2D comparison to be the better approach.

Multiple Images at Once. Our goal is to evaluate the realism of *whole* procedural rules, not individual models (or models created by artists), see "Procedural Aspect" above. For such an evaluation, we have to consider not only the quality of the individual models, but also the amount of feature variance between the models, see Figure 1. Similarly to [WFC*09], we therefore allow the participants to see all the images at the same time, in our case in a 6×2 grid design, showing 12 images at once. The number of images was chosen as trade-off between screen size, how well the images, in our opinion, represented the dataset, and expected participant fatigue. Refer to Section 7 for further discussion. The layout of our experiment's screen can be seen in Figure 2.

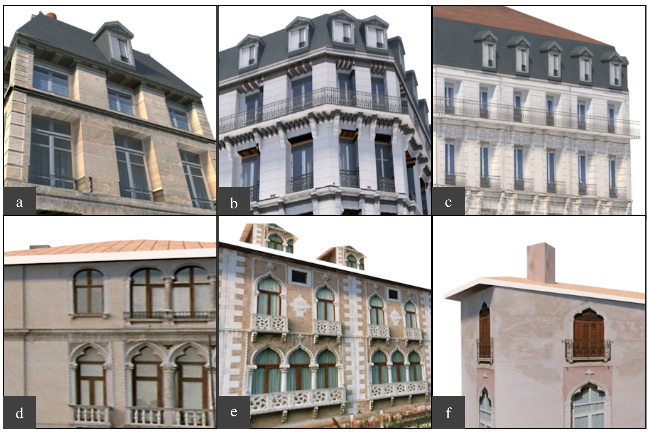
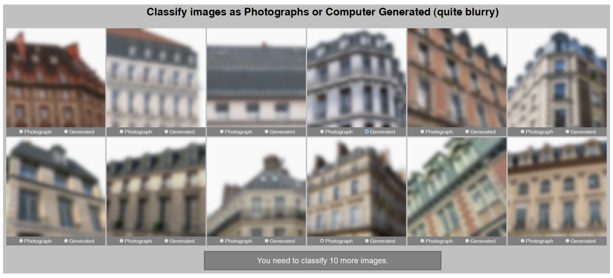


Figure 1: *Shared features and lack of variation in procedural models.* Top row*: three out of six procedurally generated images of Paris, while different in style, share the same roof windows.* Bottom row*: Three Venetian buildings sharing the same flaw, a white rim around the roof.*



Classify images as Photographs or Computer Generated (quite blurry)

You need to classify 10 more images.

Figure 2: *Classification screen, shown for the Paris dataset at 37px Gaussian blur, with 1 image classified as computer generated, 1 classified as a photograph, and 10 images left unclassified.*

Structure and Fine Detail. It is commonly alluded that realism is especially influenced by the presence or absence of fine detail [REKS13, JDA*11, GGG*16]. Since the layout of a procedural model is often stochastically determined, we also wished to investigate the role of the coarser features and the structure of a building.

To do this, we assume a roughly similar feature size across images and apply a Gaussian blur to filter out the higher-frequency detail at six different blur levels: 1px (identity – no blur), 7px, 13px, 25px, 37px, and 55px ($\sigma = 0.3(\frac{\text{blur size in px}-1}{2} - 1) + 0.8$), see Figure 3. These values are based on the results of a pilot study and range from a small blur (7px) that only removes very fine detail to a blur level (55px) that some considered impossible, see Section 5.2, but that still preserves information about realism, see Section 5.1.

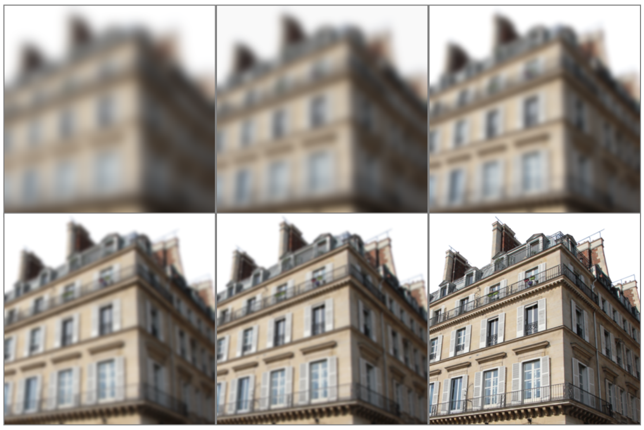


Figure 3: *The 6 different Gaussian blurs, applied to a Paris dataset image, see below. Left to right, top to bottom: 55, 37, 25, 13, 7, and 1px (identity) blur.*

The assumption about uniform feature size is a strong one. However, an inspection of the images used in the experiment (see Supplemental Material) shows that the size of features, e.g. windows and floors, varies only slightly between images. Additionally, sizes of features in both photographs and procedural models are not easily available. Our approach is not ideal and does introduce a bias, but we consider it an acceptable compromise.

Datasets. For our experiment, we have used four *pre-existing* architectural procedural rules, namely Favela, Medieval, Paris, and Venice, see Figure 4. These procedural rules were created by advanced practitioners, matching the profile of the vast majority of procedural rule authors. For each procedural rule, we generated the corresponding images and gathered photographs, as described in Section 4, creating four *datasets*. The procedural rules used in our experiment have differing characteristics, see Table 1, and provided us with a reasonable range for the experiment. Unfortunately, we were unable to acquire any procedural rules that would generate other kinds of buildings, such as family houses, industrial or commercial buildings.

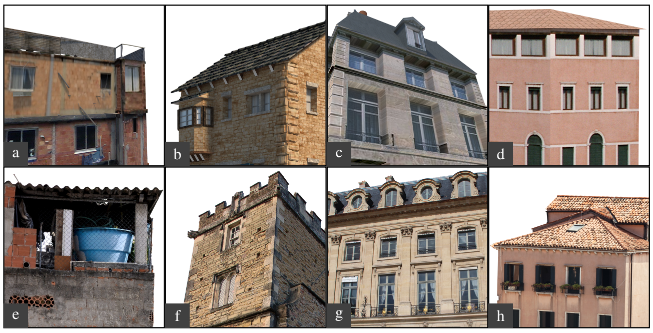


Figure 4: *Sample images from our datasets. Left to right: Favela, Medieval, Paris, Venice. Top row are generated images, bottom row are photograph cut-outs.*

Quality	Textures	Geometry	Overall
Favela	Very High	High	Very High
Medieval	Medium	Medium	Low/Medium
Paris	High	Medium	Very High
Venice	Low	Low	Low

Table 1: *Subjective pre-experiment assessment of procedural rule quality.*

4. Method

In this section, we first describe the dataset creation process and then describe the experiment itself.

4.1. Stimuli – Image Sets

For each dataset, we need to create two sets of images – one set of photographs and one set of generated images – to present to participants for comparison. Each of these two image sets needs to a) be representative of the underlying data, and b) provide the smallest possible number of clues as to the origin of the images. There is a trade-off between these two qualities and human judgement is required.

Choosing Photographs. If the photographs that were used in creating the procedural rule were available, we could just reuse them. However, in our case original photographs are not available, therefore we have to collect and process a new set. We chose publicly available photographs with permissive licenses (see Supplemental Material) and conjectured that people tend to photograph (and know) an object from the most representative viewpoints.

Furthermore, we selected our photographs to match the range of models the respective procedural rules typically produce.

Next, we removed photographs with too many occlusions from objects such as cars and people. There were two exceptions to the rule. First, for some sets of buildings, it is practically impossible to find images without people in them. Therefore, those images cannot be excluded. We deal with this issue later. Second, we have not excluded photographs with artifacts that could plausibly be a part of the procedural model, such as wall-mounted street signs or lamps; ideally, they would be generated by the procedural rule. Finally, we remove images with *extreme* angles, exposures, graining, or which are otherwise severely defective.

Generating CG Images. We roughly match the camera angles (and focal lengths) between the above set of photographs and the set of generated images we are creating. This manual process seems to be relatively easy to execute in practice and makes sure we do not introduce any *new* bias into the experiment. Nonetheless, our preliminary experiments had shown that people are not good at judging the plausibility of camera angles for photographs, nor computer generated images. For that reason, we have instructed participants not to pay attention to the camera angles. Refer to Section 5.2 for further discussion.

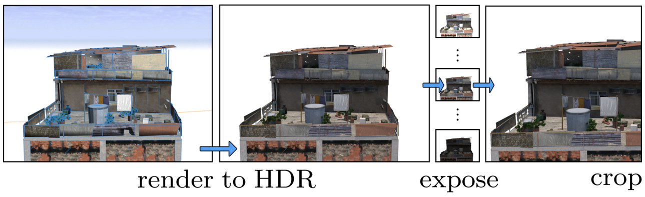


Figure 5: *Computer generated image preparation workflow. A 3D model is rendered to an HDR file with full global illumination, the exposure is manually adjusted, and the image is cropped.*

First, we generated a number of 3D models from each rule; we then manually positioned the camera for each model to roughly match one of the reference photographs. For each image, we set the sun's elevation to a random angle between 0 and 60 degrees and the sun's azimuth to be behind the camera, with a deviation of ±60 degrees. Next, we rendered our images against a white background, at a resolution of 4000×3000 pixels, until convergence. We used a commercial renderer with full global illumination and Hošek-Wilkie sky model [HW12] support. A white ground plane has been used to approximate light interactions with the environment. Finally, we manually adjusted each image's exposure to prevent over- and under-exposure. This process, together with the cropping step discussed below, is shown in Figure 5.

Image Pre-processing. At this stage, we have a set of photographs and a set of computer generated images ready. Before they can be used in the experiment, both sets need to be pre-processed further.

First, for both photographs and generated images, we always attempt to isolate single buildings, even if they stand in a row of other buildings. This is done to present the participant with only one building in each image and remove bias from the experiment. To this end, we crop the images to a square so that regions with people or other occluders can be avoided, and all images have an identical aspect ratio.

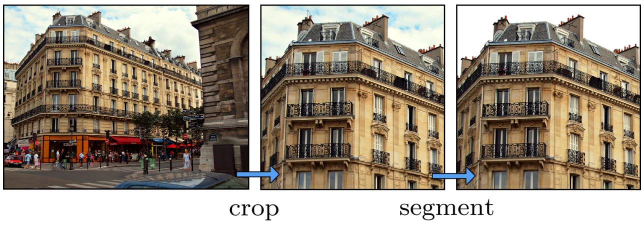


Figure 6: *Photograph preparation workflow. A photograph is first cropped so that people and other biases are removed. It is then segmented to remove the background. The cropping process is nearly identical for the computer generated images.*

Second, we remove the background from our *photographs* and replace it with a solid white color, see Figure 6. This is a crucial concept in our experiment, as it eliminates a large quantity of bias. For buildings, segmentation is a relatively easy task, as the majority of the edges are straight.

While putting photograph cut-outs against a white background will introduce some bias, alternatives such as replacing all backgrounds with a procedurally generated scene or a collage of a photograph and a rendering would introduce an even stronger bias. We think the white background approach is the least objectionable one. Additionally, laymen are familiar with photo cut-outs as they are a popular desktop publishing technique, frequently used in magazines and textbooks.

Finally, we resize all images to a common size of 300×300px. At this resolution, reasonable detail can be discerned on both photographs and generated images without making it prohibitively difficult to prepare high quality data.

Another source of bias that could possibly be dealt with is the color-tone of the images. However, with the exception of removing extremely discolored images, which we removed in a previous step, we do not adjust color-tone over our datasets.

Finally, to prevent ourselves from unconsciously introducing further bias by choosing and/or crafting the best/worst images for our experiment, we created a larger amount of data than we needed: 21, 29, 49, and 33 candidate photographs and 24, 32, 34, and 33 candidate computer generated images for the Favela, Medieval, Paris, and Venice datasets respectively. From those, 6 computer generated images and 6 photographs were randomly chosen for each dataset.

4.2. Experiment

We now describe the experiment itself. An overview of the procedure is given in Algorithm 1. Our experiment was conducted in person; however, the methodology could easily be used online.

The participants started by filling out a personal questionnaire and familiarizing themselves with the UI on a training screen. The individual classification screens (described below) followed. Finally, the participants were debriefed.

To reduce variance, we randomly chose 6 generated images and 6 photographs from each dataset once, and reused them for *all* participants. We started from the strongest blurs, hypothesizing that

Algorithm 1: Experiment structure as pseudo-code.

```
for dataset in datasets do
    dataset = RANDOM.PICK6CG6REAL(dataset);
blurs = LIST(55px, 37px, 25px, 13px, 7px, 1px);
for each participant do
    SHOWPERSONALQUESTIONNAIRE();
    SHOWTRAININGSCREEN();
    datasets = RANDOM.PERMUTE(datasets);
    for blur in blurs do
        for dataset in datasets do
            images = APPLYBLUR(blur, dataset);
            images = RANDOM.PERMUTE(images);
            SHOWCLASSIFICATIONSCREEN(images);
    SHOWDEBRIEFINGSCREEN();
```

this minimizes the learning effect, as the participants cannot be sure about their previous – "more blurry" – choice. We interleaved the datasets to further minimize how well participants recognize and remember images and their associated classification choices. Finally, we randomized the position of images on the 6×2 grid and the order in which the datasets were interleaved to prevent spatial and temporal biases.

Classification Screen. On the classification screen the participants choose an *assigned class* for each image: either *generated* or *photograph*. They are unaware of the image's *original class*. The classification screen, depicted in Figure 2, showed 12 images presented in a 6×2 matrix against a gray background. The participants were instructed to take as much time as they needed, and told that usually the test takes around 20 minutes, or under a minute per screen. The participants then used a radio button to pick an assigned class for each image. Once all images were classified, the participant would confirm the selection using a button. A slideshow walkthrough of the whole experiment can be found in the Supplemental Material.

Participants. In total, 52 participants, mostly university students, on average 24.33 years old ($SD = 4.83, \min = 20$, $\max = 43$), of which 11 were female, took part in our experiment. All had normal or corrected to normal vision, and normal color vision. Reported experience with related fields is given in Table 2.

Experience	None	Casual Interest	Professional
Comp. Graphics	27	23	2
Photography	22	30	0
Architecture	41	11	0

Table 2: *Experience in related fields as reported by the participants.*

5. Data Analysis

Below, we analyse the quantitative (classification) and qualitative (questionnaire) data gathered in our experiment. Source code for our quantitative analysis, as well as additional analysis data, is provided in the Supplemental Material.

5.1. Quantitative Analysis

One of the main goals of our study is to investigate the role of various scales (levels) of detail on perceived generated building realism. Specifically, we postulate the two following hypotheses: H_{DIFF}: participants can tell photographs and generated buildings apart, and H_{SCALE}: the detail that allows participants to tell photographs and generated buildings apart is present at various scales. We test these hypotheses by studying the classification accuracy of building images subjected to different levels of blur. Since our hypotheses only concern one-way interactions, we limit the post-hoc tests reported here to the main effects. The rest of the analysis is exploratory. We therefore only present descriptive statistics for the two and three-way interactions.

Hypothesis H_{DIFF}. To test hypothesis H_{DIFF}, we use the Student's t-test to compare mean classification accuracy at the highest blur level, 55px, and random choice (= 0.5). We define *accuracy* as the fraction of classifications in which the assigned class equals the original class. The difference for the overall accuracy at a 55px blur is statistically significant ($t(2495) = 10.2, p < .001$). This demonstrates statistically significant differences in the images that allow the participants to distinguish between computer generated images and photographs, even at the highest blur level. This holds for each individual dataset with the exception of Favela, for which the accuracy at a 55px blur is not different from $\mu = 0.5$ in a statistically significant way.

A weaker, but more relatable variant is that there is a statistically significant difference between photographs and generated images at no blur ($t(2495) = 42.25, p < .001$). This means that users can differentiate between photographs and generated images without blur. For this variant, the statement holds for all individual datasets.

Hypothesis H_{SCALE} and ANOVA. To test the hypothesis H_{SCALE} and further explore the data, we use three-way, within-subject, ANOVA [Bak05] to model differences in classification accuracy (the dependent variable) with blur level, dataset, and an image's original class as factors (the independent variables). There are six blur levels (1px, 7px, 13px, 25px, 37px, 55px), four datasets (Favela, Medieval, Paris, Venice), and two original classes (computer generated, photograph).

Blur level ($F(5,255) = 93.3$, $p < .001$) and dataset ($F(3,153) = 82.6$, $p < .001$) are the main statistically significant effects impacting accuracy. In other words, classification accuracy is significantly better for some blur levels than for others. Similarly for datasets. The image's original class (*computer generated* or *photograph*) does not have a significant effect on accuracy ($F(1,51) = 2.57$, $p = .11$). The relationships between the significant effects and classification accuracy are shown in Figure 7.

Post-hoc tests (Tukey's HSD for each factor individually) reveal significant differences between all blur levels ($p < .001$, for 7px and 13px, $p = .014$) with the exception of 25px and 37px ($p = 1.0$), 25px and 55px ($p = .15$) and 37px and 55px ($p = .21$). For the dataset effect, the post-hoc tests show that the difference in accuracy is only significant between the Medieval and the Favela datasets ($p = .009$). Other differences are not significant. Therefore, we can conclude that there isn't enough difference

Dataset	All		Favela		Medieval		Paris		Venice	
Assigned Class ⇒	CG	Photo	CG	Photo	CG	Photo	CG	Photo	CG	Photo
Original Class: CG	79.89%	20.11%	71.47%	28.53%	95.19%	4.81%	69.55%	30.45%	83.33%	16.67%
Original Class: Photo	15.30%	84.70%	26.28%	73.72%	4.81%	95.19%	17.31%	82.69%	12.82%	87.18%

Table 3: *Contingency tables for 1px blur. For all combinations of original image classes and datasets, we give the percentage of images that are assigned the classification computer generated (CG) and photograph (Photo). The equal values for the Medieval dataset are not an error.*

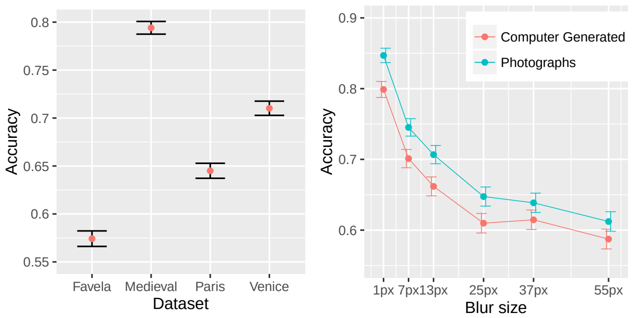


Figure 7: Left*: Average classification accuracy for each dataset;* Right*: Average classification for generated images and for photographs. The accuracy plots show the relationship between blur level and accuracy for both original classes. The similar shape and position of plots illustrates the lack of a statistically significant effect of an image's original class. For both plots, the error bars show the standard error.*

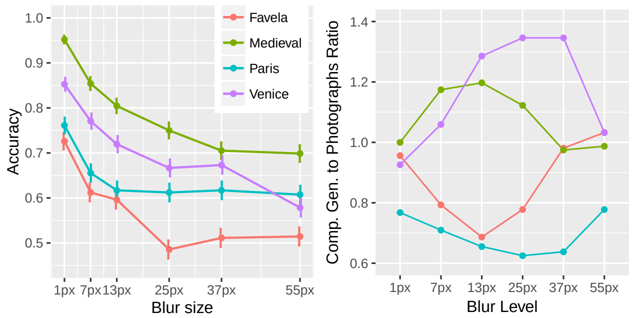


Figure 8: Left*: The accuracy of each dataset across blurs. Notice the dip at 25px for the Favela dataset and the decrease in accuracy between 37px and 55px for the Venice dataset.* Right*: (same legend as Left). For each blur level and dataset, the ratio of images classified as computer generated and images classified as photographs. If the ratio is greater than* 1*, more images were classified as computer generated than as photographs. If the ratio is smaller than* 1*, more images were classified as photographs than as computer generated.*

between these blur levels for them to be considered different. However, it does not imply that no further features would be removed by more extreme blurs or that there is no change between these blurs, only that the change is not statistically significant.

Because the effect of blur size is statistically significant, hypothesis H_{SCALE} can be accepted. Below, we provide further exploration of the data.

Two- and Three-way interactions. There are two significant two-way interactions. The first is between images' original class and dataset ($F(3,153) = 21$, $p < .001$). This signifies that the difference in accuracy between computer generated images and photographs changes between datasets. This is shown in Table 3 which shows the percentages with which each original class is classified as either a photograph or computer generated. While it is clear that the classification percentages differ, it is not clear whether this is caused by a different quality of photographs or a different quality of generated images. We speculate that the quality of photographs (lens, composition, experiment pre-processing, etc.) could be assumed to be *approximately* equal or at least have less variance than the quality of the generated images. We therefore stipulate that it is mainly the quality of the generated images and therefore the individual datasets that changes.

The second significant two-way interaction is between blur level and dataset ($F(15,765) = 3.96$, $p < .001$). This implies that the blur behaves differently for each dataset. We speculate that this result is most likely an artifact. As visualized in Figure 8 (Left), the differences in behaviour manifest themselves in two places: in the Favela dataset at a 25px blur, where it is likely the result of insufficient sample size, and at 55px for the Venice dataset. For the latter, one plausible explanation is that the Venice dataset has less detail at that resolution than the other datasets.

The third two-way interaction, between the image's original class and the blur, is not significant ($F(5,255) = 0.34$, $p = .89$). See also Figure 7.

Finally, the three-way interaction between all factors is significant ($F(15,765) = 6.49$, $p < .001$). Figure 9 illustrates this; blur behaves differently between computer generated images and photographs for each dataset.

Computer Generated to Photographs Ratio. For each dataset, Table 3 shows the contingency table for the image classes. As mentioned above, our experiment's design does not allow us to verify whether the differences in the percentages between datasets are solely due to the quality of the generated images, or whether the quality of photographs plays a role.

Figure 8 (Right) shows the ratio of the number of images classified as computer generated versus the number of images classified as photographs. Both the Medieval and Venice datasets have a number of blur levels for which the number of images classified as computer generated is higher than the number of those classified as photographs (ratio > 1). In other words, photographs are being classified as computer generated. For the Medieval dataset,

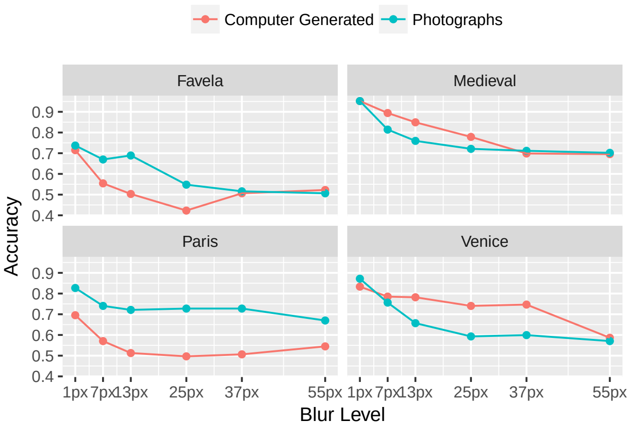


Figure 9: *Three-way interaction between original class, dataset, and blur level. For each dataset, the way blur behaves with respect to original class and dataset differs.*

this might be due to the relative obscurity and perhaps implausibility, from today's perspective, of such buildings. For the Venice dataset, the fact that many of the photographs were taken almost perpendicularly to the facade might have played a role. However, no conclusive evidence is available.

For the Paris dataset, the participants overestimated the realism of the images throughout the experiment. Interestingly, they thought the dataset contained more photographs in roughly the middle of the experiment (37-13px) than at the beginning and at the end. A similar, more pronounced effect, was recorded for the Favela dataset.

Overall, datasets that we assessed as "more realistic" before the experiment (Paris, Favela), see Table 1, better convince participants of their realism (ratio < 1, meaning some of the generated images were classified as photographs) even when shown at higher blurs.

Summary. We accept the premise of hypothesis H_{DIFF}: that *overall* participants perform better than random choice even at the highest blur level ($p < .001$). A three-way, within-subject ANOVA was used to fit a model to our data; because ANOVA shows that the blur level has a statistically significant effect on classification accuracy ($p < .001$), we can accept hypothesis H_{SCALE}. We therefore conclude that there exist certain image features that are removed or diminished by each additional blur level that contribute to the perception of realism – up to and including the 25px blur, as shown by the post-hoc tests. Higher blurs have not been shown by the post-hoc tests to have a statistically significant impact on accuracy.

Finally, we have conducted further exploratory analysis and discussed the other statistically significant effects in our experiment.

5.2. Qualitative analysis

In this subsection, we detail the participants' debriefings, analyse the outlier images, and discuss individual datasets.

Debriefing. After the test, each participant is debriefed (see the Supplemental Material). First, we enquired about the factors that influenced their decisions. We then asked if there were any telltale signs, or *indicators*, which influenced their classification of the images.

By manually tallying the number of times certain topics were discussed in the debriefings, we identified 9 indicators that influenced the 52 participants' decisions the most, namely: imperfections and/or small detail (30 participants $\approx$ 58%), texture (19 $\approx$ 37%), reflections in windows (18 $\approx$ 35%), "weird" or uniform color (17 $\approx$ 33%), things in, behind, or around windows (16 $\approx$ 31%), model structure (14 $\approx$ 27%), lighting (12 $\approx$ 23%), shadow (12 $\approx$ 23%), and regularity (11 $\approx$ 21%).

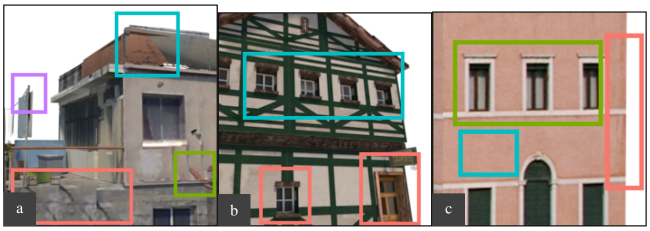


Figure 10: *Frequently recognized computer generated images with artifacts (not to scale). (a) Texture repetition (red), incorrect geometry (purple, green, cyan); (accuracy* $a = 0.88$*) (b) model structure (red), regularity and window detail (cyan); (*$a = 0.96$*) (c) linear edges, see below (red), lack of window reflections (green), uniform color/lack of detail (cyan), (*$a = 0.98$*).*

Overall, participants looked for, and saw, imperfections and small detail (stains, dirt, wall cracks, etc.) as indicators of realism. On the other hand, the lack of detail, for example large areas of uniform color (Figure 13c), was considered unrealistic, see Figure 10.

Texture quality, especially inadequate resolution, inadequate detail, or incorrect repetitions (for example, the bottom row of Figure 1) were also identified as indicators of a generated model. Some participants also mentioned such terms as "plasticity", referring to the overuse of bump-mapping, see Figure 1e.

Reflections in windows were another important factor in perceived realism. The consistency of reflections, the curtain position, and the interiors were all indicators discussed by participants. However, analysis of the images suggests that when participants though they saw a real reflection in a photograph, they really just saw a texture on the window's model.

Scene lighting, color, and shadows were another often indicated factor. A subset of the 17 participants ($\approx$ 33%) that mentioned color remarked that a real building couldn't have such colors, one remarking that the colors were "kitchy". Some considered the lighting overall to be unrealistic, often mentioning shadow sharpness as an indicator of a computer generated image. See [RLCW01] for further discussion of how shadow sharpness influences perception of realism, possibly misguiding participants. Additional participants noticed apparent overexposure and light bleeding around the edges of the photographs.

Finally, model structure and regularity had an effect. Participants were sensitive to repeated identical instances of windows (e.g., top row of Figure 1), irregularities, and were sometimes uncertain

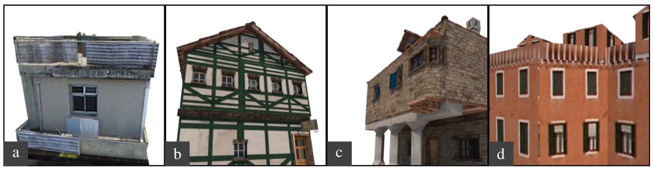


Figure 11: *Computer generated images with implausible structures. (a) Balcony and roof without any access points (*$a = 0.87$*); (b) Window placed in a support pillar, door too close to edge (*$a = 0.96$*); (c) Wooden balcony support inside stone/concrete support (*$a = 1.0$*) (d) Deformed roof structure and incorrectly applied texture (*$a = 1.0$*).*

whether such a building was structurally plausible (e.g., Figure 10 and Figure 11). However, we did not observe participants noticing repetitions between models, such as those in Figure 1, beyond the above mentioned instance.

Cut-Out Edges, Camera Angles, Background. During debriefing, we asked the participants whether the edge between the building and the white background influenced their choices. Of the 52 participants, 21 (≈ 40%) remarked that the cut-out edge played a role. Some mentioned that generated images had a sharper edge; this, however, is unlikely considering the generated images were significantly downscaled after segmentation. A small number mentioned that the complexity of the silhouette influenced their decisions. Notably, the sharpness and fine detail of edges are lost even at a low (7px) blur level, meaning these observations relate mostly to images without blur.

During debriefing, 8 participants (≈ 15%) said that the white background made them choose computer generated more often. This clearly biased our experiment. We believe this bias is hard to avoid, see discussion in Section 4.1.

In a pilot experiment, we observed that participants are not good at determining the realism of camera angles. We therefore instructed participants not to consider camera angles. After the experiment, we asked the participants whether the camera angles influenced their decision, with 18 participants (≈ 35%) admitting they have been influenced for at least one image.

Other Factors. The majority of the 52 participants found the higher blur levels quite challenging and some found it almost impossible. Many commented that the experiment was easy or not too difficult for the lower blur levels. Two participants (≈ 4%) considered the experiment too long and tiring.

Five people commented that they were or might had been influenced by their previous image choices. We suspect this had a small or negligible overall effect, as accuracies of individual images changed steadily between blurs.

Images. Some of the factors that affect realism were already discussed above and illustrated in Figure 10. Here, we take a further look at some individual images and blur levels.

Overall, at high blur levels (Figure 12) participants seemed to look for two factors: plausible building structure and irregularity.

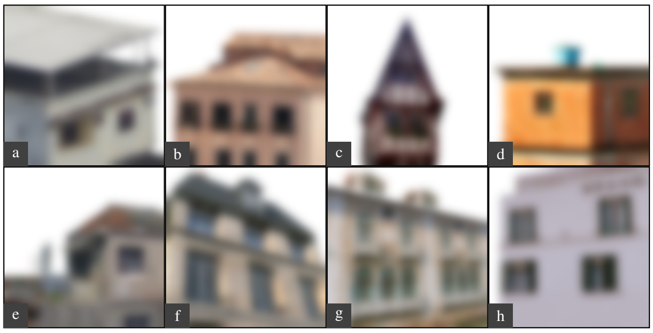


Figure 12: *Most confounding buildings at the highest, 55px level blur.* Top row*: photographs that were classified as computer generated images the most. With fine detail missing, participants had to rely on coarser structure to make their judgements. Note how the structure seems implausible (a,b,c), the building improbable (c), and the color too strong (d).* Bottom row*: computer generated images that were classified as photographs the most. The plausible structure (e, f, g) contributes to realism. For (h), the variance in the window shutter positions and the fine detail in the top right corner are the most likely factors.*

Sometimes, these indicators can be misleading, e.g., Figure 12e was thought to be very realistic at a 55px blur ($a = 0.25$), but was readily recognized as computer generated ($a = 0.88$) when shown without blur (1px blur, shown in Figure 10a). Other times, these factors can diminish realism, e.g. a photograph (Figure 12a) with a relatively high amount of detail (fine structures, stained roof, very irregular) only had an accuracy of $a = 0.23$ at the highest blur, and remained on par with random choice ($a = 0.5$) even when blur was removed (Figure 13b).

For images without blur, fine detail, or lack thereof, becomes an additional factor. However, a high level of fine detail can be overridden by other factors, see Figure 13.

Overall, realism is a multi-faceted problem where a qualitative study might provide guidelines, but cannot be relied upon to easily predict performance.

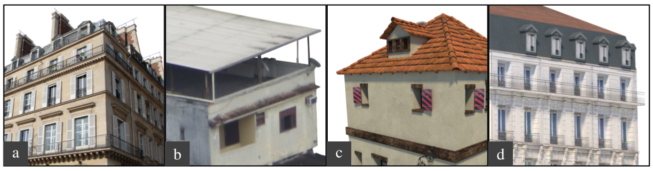


Figure 13: *Unblurred (1px blur) images. (a,b) Photographs that were often classified as computer generated (*$a = 0.67$ *for (a) and* $a = 0.5$ *for (b)). For (a), the high regularity of the unusual chimneys might be the cause. For (b), the implausible layout of the building is the probable cause. (c, d) Computer generated images that were very often classified as such (*$a = 1.0$ *and* $a = 0.98$*, respectively). For (c), the simple structure, uniform walls, and lack of detail on both edges and surface, and the very strong roof color seem to be the decisive factors. For (d), the large quantity of repetition (windows, roof windows), the unusual roof shape, and perhaps lighting are the most likely factors diminishing realism.*

Participants. The participants were mostly university students, some with casual interest in photography, computer graphics, and architecture, see Table 2. Therefore, our study best represents a non-expert audience similar to the one that is the final consumer of procedurally generated buildings. As corrobated by [GMM15], there might be differences in how expert and non-expert participants perceive realism. It is therefore important to match study participants with the intended audience. Refer to the Supplemental Material for additional statistics about the participants.

Summary. We have collected and discussed factors that participants subjectively considered most important for the perception of realism, and illustrated them with images from the experiment. Additionally, we have analysed individual images with both subjectively and objectively interesting features, illustrating the complexity of the interplay between the above factors.

6. Results

In this paper, we have presented the results of a study that investigates the perceived realism of procedural buildings. Our study has a quantitative and a qualitative part.

In the quantitative part, we have shown that even at the highest level of blur, participants on the whole still perform above random choice (H_{DIFF}), and that blur has a statistically significant effect on the resulting accuracy (H_{SCALE}). This shows that realism in procedural models is carried not only by fine detail, but also by coarser detail and structure at various scales.

The qualitative part has indicated that quality of textures, colors, presence of reflections, presence of model irregularity, structural plausibility (Figure 10a, b), and lack of silhouette detail (Figure 10c) were important factors. Of the above, irregularity, plausibility, and quality of texture mapping are especially relevant to procedurally generated buildings.

Contrary to our expectations, participants seemed mostly unable to spot repeated features shared between different buildings from the same dataset, such as those in Figure 1. We suspect this is due to the relatively low sampling power of our experiment, as only six generated images are shown. A different design would help to answer this question.

Lastly, we have analysed the individual images. The findings are in agreement with the above and show that realism needs to be achieved on both fine and coarse scales.

Our experiment tried to answer a question in a difficult context, where any possible design would have a number of biases, such as the rendering method, cut-out quality, lighting, or selection of data. We attempted to minimize the number and effect of these biases. The biggest biases in our study were the effect of the cut-out background, lighting control, and a slight memory effect. Those biases, however, are not strong enough to invalidate the results of our study, see Section 5.2 and Section 3 for an analysis.

Below, we discuss the applicability of our results and summarize them into recommendations.

Applicability. Among other information, our study of procedural buildings gives us insights into how well procedural rules deal with structure of buildings, feature repetitions across instances, etc. Such insights could not be revealed by studying manually reviewed or modelled buildings, where conceivably, the artist would remove many of the imperfections. A study of that kind would only provide data about the influences of different rendering factors or errors that were intentionally included in the model.

In that sense, we surmise our hypotheses H_{DIFF} and H_{SCALE} might provide us with results that differ from results one would get from a study of manually modelled buildings, as we do not avoid or choose our errors, but instead sample them from procedural rules to identify what they are.

Consequently, our results are most readily applied to procedural buildings. However, they are also at least partially applicable to non-procedural buildings. Both desirability of various details and structural plausibility can be considered sensible guidelines for any virtual buildings. Others results, especially those related to shadows, colors, and other rendering aspects, can be applied to non-procedural buildings without caveats.

Recommendations. On the whole, participants could differentiate between generated and photographic images (H_{DIFF}). The quantitative analysis shows that there is information that contributes to the participants' perception of realism even if the small and very small detail gets blurred out (H_{SCALE}). Authors should therefore focus not only on this detail, but also on what is left when the small and very small detail is taken away – the overall structure.

The debriefings suggest that both manually and procedurally modelled buildings should provide enough content to convince their users that the building interacts with the real world. Such content could include dirt, weathering or cracks, etc. Another, relatively cheap, way to boost realism is by adding reflections to windows, either from a texture, or from an environment map.

Additionally, procedurally generated buildings seem to suffer the most from poor structural plausibility and too much regularity. Users notice misaligned, misplaced, or repetitive geometry and texture, both for smaller regions as well as for the overall structure. A good procedural model should strive to provide plausible structure and placement of its components and textures at all scales.

7. Limitations and Future Work

With this study we want to stimulate further research into the realism of procedural modeling in general. We hope our results will motivate research into procedural rules with improved plausibility and correctness. For example, the ability to generate meaningful interiors for a given exterior, creating procedural models that are plausible from a navigation and plumbing perspective, or a system that is able to detect and reject implausible, or poor quality outputs.

As for realism itself, we hope to see work that would study procedurally generated buildings and their perception in context of other buildings; as Ramanarayanan has shown [RBF08], there could be several further effects at play. Such an experiment could also help with better answering the question of how much participants notice and are influenced by common features and regularity in a set of procedural buildings.

Additionally, the recent advent of easily accessible image classification using neural networks [SIV16] could open doors to novel algorithms that replace the human participants in our study with artificial participants. This would allow procedural model synthesis to be guided by a learnt concept of realism.

Our technique could be adapted and applied to large scale urban models from any source, including those which were manually created, scanned, reconstructed, or found in large databases. The only general requirement is that a photographic ground-truth is available. It may also be possible to apply the technique to other domains than urban environments.

As in all user studies, the main limitation of our work is that our results are, in the strictest sense, valid only for our datasets and our setup. We realize other kinds of datasets, such as family houses, industrial buildings, commercial buildings, etc., might behave differently and that we have only explored a small fraction of the buildings even our datasets can generate. Other setups, where the buildings are seen in aggregates, in motion, or under different conditions could provide different insights.

Acknowledgements

We would like to thank the anonymous reviewers for their valuable input. The study was funded by ESRI, the Charles University in Prague's Grant SVV-2016-260332, and the Czech Science Foundation grant 16-18964S. Filip Děchtěrenko was supported by Czech Academy of Sciences Strategy AV21 – RVO 68081740. The computer generated images were rendered using the Corona Renderer, kindly donated by Render Legion. The Favela dataset was graciously provided by Matthias Bühler of vrbn.io.

References

[Bak05] Bakeman R.: Recommended effect size statistics for repeated measures designs. *Behavior research methods 37*, 3 (2005), 379–384. 5

[Čad08] Čadík M.: *Perceptually Based Image Quality Assessment and Image Transformations*. PhD thesis, CTU Prague, January 2008. 2

[Dal92] Daly S. J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE/IS&T 1992* (1992), International Society for Optics and Photonics, pp. 2–15. 2

[DHL*98] Deussen O., Hanrahan P., Lintermann B., Měch R., Pharr M., Prusinkiewicz P.: Realistic modeling and rendering of plant ecosystems. In *Proc. SIGGRAPH* (1998), ACM, pp. 275–286. 1

[DSG*12] Doersch C., Singh S., Gupta A., Sivic J., Efros A. A.: What makes Paris look like Paris? *Proc. SIGGRAPH '12 31*, 4 (2012). 2

[Ebe03] Ebert D. S.: *Texturing & modeling: a procedural approach*. Morgan Kaufmann, 2003. 1

[FW16] Fan L., Wonka P.: A probabilistic model for exteriors of residential buildings. *ACM Transactions on Graphics (TOG) 35*, 5 (2016), 155. 1

[GGG*16] Guérin E., Galin E., Grosbellet F., Peytavie A., Geneveaux J.-D.: Efficient modeling of entangled details for natural scenes. *Computer Graphics Forum (Proceedings of Pacific Graphics 2016) 35*, 7 (2016). 3

[GMM15] Gain J., Merry B., Marais P.: Parallel, realistic and controllable terrain synthesis. In *Computer Graphics Forum* (2015), vol. 34, pp. 105–116. 9

[HW12] Hošek L., Wilkie A.: An analytic model for full spectral skydome radiance. *ACM Transactions on Graphics 31*, 4 (2012), 95. 4

[JDA*11] Johnson M. K., Dale K., Avidan S., Pfister H., Freeman W. T., Matusik W.: CG2Real: Improving the Realism of Computer Generated Images using a Large Collection of Photographs. *IEEE TVCG 17*, 9 (2011), 1273–1285. 1, 3

[KCKK12] Kalogerakis E., Chaudhuri S., Koller D., Koltun V.: A probabilistic model for component-based shape synthesis. *ACM Trans. Graph. 31*, 4 (July 2012), 55:1–55:11. 1

[MMW*11] Mathias M., Martinovic A., Weissenberg J., Haegler S., Van Gool L.: Automatic architectural style recognition. *ISPRS XXXVIII-5/W16* (2011), 171–176. 2

[MWH*06] Müller P., Wonka P., Haegler S., Ulmer A., Van Gool L.: *Procedural modeling of buildings*, vol. 25. ACM, 2006. 1, 2

[NGDA*16] Nishida G., Garcia-Dorado I., Aliaga D., Benes B., Bousseau A.: Interactive sketching of urban procedural models. *ACM Trans. Graph.* (2016). 1

[RBF08] Ramanarayanan G., Bala K., Ferwerda J. A.: Perception of complex aggregates. In *ACM TOG* (2008), vol. 27, ACM, p. 60. 2, 9

[REKS13] Reinhard E., Efros A. A., Kautz J., Seidel H.-P.: On visual realism of synthesized imagery. *Proceedings of the IEEE 101*, 9 (2013), 1998–2007. 1, 3

[RFWB07] Ramanarayanan G., Ferwerda J., Walter B., Bala K.: Visual equivalence: towards a new standard for image fidelity. In *ACM TOG* (2007), vol. 26, ACM, p. 76. 2

[RLCW01] Rademacher P., Lengyel J., Cutrell E., Whitted T.: Measuring the perception of visual realism in images. In *Rend. Techn.* Springer, 2001, pp. 235–247. 1, 2, 7

[S*80] Stiny G., et al.: Introduction to shape and shape grammars. *Environment and planning B 7*, 3 (1980), 343–351. 1

[SIV16] Szegedy C., Ioffe S., Vanhoucke V.: Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR abs/1602.07261* (2016). 10

[SM15] Schwarz M., Müller P.: Advanced procedural modeling of architecture. *ACM Transactions on Graphics 34*, 4 (Proceedings of SIGGRAPH 2015) (Aug. 2015), 107:1–107:12. 1, 2

[TYK*12] Talton J., Yang L., Kumar R., Lim M., Goodman N., Měch R.: Learning design patterns with bayesian grammar induction. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2012), UIST '12, ACM, pp. 63–74. 1

[WBSS04] Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing 13*, 4 (2004), 600–612. 2

[WFC*09] Wallraven C., Fleming R., Cunningham D., Rigau J., Feixas M., Sbert M.: Categorizing art: Comparing humans and computers. *Comp.&Graph. 33*, 4 (2009), 484–495. 2

[WWSR03] Wonka P., Wimmer M., Sillion F., Ribarsky W.: *Instant architecture*, vol. 22. ACM, 2003. 1

FAQ for On Realism of Architectural Procedural Models
Version 1, 14. May 2017

Q: What system did you use to generate the buildings?
A: We used CityEngine, a state of the art system for generating procedural buildings.

Q: What system did you use to render the buildings?
A: We used the Corona Renderer, a non-biased, fully GI enabled renderer.

Q: Where do your datasets come from?
A: The Venice, Paris, and Medieval datasets are part of CityEngine's examples. The Favela dataset has been gracefully provided by Matt Buehler of VRBN.io.

Q: Is your paper just benchmarking the CityEngine results?
A: For the most part, the answers is no. We do indeed use CityEngine, and 3 out of 4 datasets come from CityEngine's examples. However, these datasets differ in quality and each of them represents a different set of possible mistakes an author of a procedural rule in any system might make (though we understand that there might be small differences implied by, e.g., how easily some of these mistakes can be made in different modelling systems). The results are therefore applicable to buildings generated by any method, within a small margin of error.

Q: Why did you not compare similar but not identical man-made models and photographs?
A: A man made model would not have the same kinds of errors a procedurally generated model, created automatically and stochastically from a man-made procedural rule, would. The errors in the generated results allow us to see what can, typically, go wrong.

Q: Why did you not compare photographs of buildings and their highly realistic man-made models?
A: Because we focus on finding errors in the models and the materials. Comparing a very well done model would only provide information about what parts of the rendering process influence realism.

Q: Why did you not use a highly realistic man-made model and turn various rendering features on and off to see how they influence the model's perceived realism?
A: See the above answer.

Q: Why did you not use a procedural rule and turn various features on and off to see how they influence the model's perceived realism?
A: Our experiment allowed us to look into errors that rule authors might introduce by accident. As such, the above is outside of the scope of our paper. However, we think such an experiment would provide the community with important results.

Q: Why did you show all of the images at once?
A: One of our hypotheses was that the users would be able to spot common patterns on procedural rules, or see some other form of systematic error that is present in the procedurally generated images, but not in the photographs. For this reason, we wanted the users to see all of the images at the same time.

Q: Why did you not use a generated and/or realistic background instead of a white one?
A: We wanted to use a background that would not give any cues as to the origin (real, generated) of the whole image. We therefore didn't want to use a generated background. The only plausible choice would be to use a real background, e.g. the photograph of a sky. However, we would then have to match the sky's intensity and light direction to the cut-out image of a building. Additionally, a realistic background might still introduce a bias towards realism. We used a solid color background which does introduce a similar bias (but towards "computer generated"), but is easier to implement.

Q: Why did you not use a different filter?
A: We wanted to use a filter that removes detail while preserving structure. We considered using a bilateral filter, a nearest neighbor filter, and a mosaicing filter, but felt those introduce a relatively strong bias towards the "computer generated" answer, as they make the resulting image look more like artwork. The Gaussian filter is akin to minification and doesn't introduce an artwork-like bias. It is also easily parameterizable.

Q: Why did you not show the image at a smaller size rather than showing them blurred?
A: We decided to show a blurred image instead of a minified image as we suspected small images might be more difficult for the users to look at. We also felt if looked at from close distance, the pixel grid of the monitor might bias the users further.

Q: Did you tell your users what the correct answers were?
A: No, we didn't tell the users the correct answers at any point in the experiment.

Q: I still do not understand the procedural part. Could you explain?
A: Our experiment investigates the error, especially in the model and its materials, a procedural rule might introduce without its author's intent. This kind of error would not be present in a manually modelled building.