STAG: Smart Tools and Applications in Graphics (2020)
S. Biasotti, R. Pintus and S. Berretti (Editors)

ViDA 3D: Towards a View-based Dataset for Aesthetic prediction on 3D models

M. Angelini[1,2], V. Ferulli[3], F. Banterle[1], M. Corsini[1], M. A. Pascali[1], P. Cignoni[1] and D. Giorgi[1]

[1]Institute of Information Science and Technologies, National Research Council of Italy
[2]Research Centre on Interactive Media, Smart Systems and Emerging Technologies - RISE Ltd, Cyprus
[3]Department of Computer Science, University of Pisa, Italy

Abstract

We present the ongoing effort to build the first benchmark dataset for aestethic prediction on 3D models. The dataset is built on top of Sketchfab, a popular platform for 3D content sharing. In our dataset, the visual 3D content is aligned with aesthetics-related metadata: each 3D model is associated with a number of snapshots taken from different camera positions, the number of times the model has been viewed in-between its upload and its retrieval, the number of likes the model got, and the tags and comments received from users. The metadata provide precious supervisory information for data-driven research on 3D visual attractiveness and preference prediction.

The paper contribution is twofold. First, we introduce an interactive platform for visualizing data about Sketchfab. We report a detailed qualitative and quantitative analysis of numerical scores (views and likes collected by 3D models) and textual information (tags and comments) for different 3D object categories. The analysis of the content of Sketchfab provided us the base for selecting a reasoned subset of annotated models. The second contribution is the first version of the ViDA 3D dataset, which contains the full set of content required for data-driven approaches to 3D aesthetic analysis.

While similar datasets are available for images, to our knowledge this is the first attempt to create a benchmark for aestethic prediction for 3D models. We believe our dataset can be a great resource to boost research on this hot and far-from-solved problem.

CCS Concepts

• ***Computing methodologies*** → ***Shape analysis;***

1. Introduction

In the past twenty years, the Computer Vision and Computer Graphics communities were united by the search for tools for the semantic analysis of content. Tasks such as image and 3D object classification, recognition, retrieval and segmentation have achieved notable performance, even comparable to humans after the resurgence of artificial intelligence techniques. Nevertheless, new sophisticated questions that go beyond content semantics are arising. Among them, important questions are: how can we predict whether people would find an image or a 3D object aesthetically pleasant? What are the visual elements that make people like some visual content, and which ones make the content unattractive? Such questions are growing in importance for many applications, including the design of tools for automatic beautification of media visual summarization, presentation of media content from large collections of images or 3D models, and understanding of social media.

Being aesthetic prediction an inherently human-centric question, learning techniques leveraging on labeled data about human preferences are expected to outperform hand-crafted solutions. Therefore, the key requirement to develop aesthetic analysis techniques is the availability of datasets featuring the visual content, along with aesthetics-related metadata to provide supervisory information. While similar dataset are available for images [MMP12], to our knowledge there is no dataset tailored to aestethic prediction on 3D models. Since a large fraction of the digital content production pipeline now involves 3D objects, we believe it timely to fill the gap and boost research on data-driven 3D visual attractiveness and preference prediction.

To this end, we describe the ongoing effort to build ViDA 3D, the first benchmark dataset in which visual 3D content is aligned with aesthetics-related metadata. The current dataset includes data about over 7700 3D models featured on Sketchfab, a popular platform for 3D content sharing. For each 3D model, we collect the number of model views occurred in-between the model upload and its retrieval, the number of likes the model got, the tags and the comments received from users. Inspired by preference prediction studies on other types of media [GUB*15], we use the metadata to associate to each 3D model two different scores as measures of its aesthetic value: the average number of likes per time span, and the ratio between the number of likes and the number of views. These measures can be used as supervisory information in learning tasks. Nevertheless, there is no consensus in the literature about how to measure aesthetic preferences of users. Therefore, additional measures can be studied and defined on the amount of information we collect, for example measures based on sentiment analysis on textual tags and comments [MMP15].

Our dataset represents 3D content as a set of 2D views: for each 3D model, we include 46 snapshots taken from different camera positions. View-based representations are popular among deep learning techniques on 3D data: 2D renderings have proven effective in describing both 3D geometry and appearance in several applications, from 3D shape classification [RROG18] to semantic segmentation [DN18]. In our context, the multiple renderings are aimed to replicate the pattern of interaction of users with 3D models before forming an opinion about their aesthetics. In other words, to mimic the process by which a human evaluates a 3D model by navigating around it and looking at its different parts. We collect 2D snapshots both concentrated around the artist's predefined position of the 3D model, and scattered across the rest of the viewing sphere. The aim is to capture information on 3D models from different perspectives, while taking into account the artists' design choices while uploading the models.

The current version of the dataset will be released for download before the end of November 2020. Subsequent releases are foreseen in a near future, after the dataset growth in size and model categories.

While many other 3D showcase platforms, such as ArtStation, cgtrader, or Quixel Megascans, could have been used as data sources, we opted for *Sketchfab* for a number of reasons. Sketchfab is a very popular platform for content sharing, with more than 3 million models organized into human-undestandable categories, and with user-generated metadata. At any rate, the major advantage with respect to other platforms is that Sketchfab models are showcased through an interactive viewer, which comes with official APIs to replicate users' patterns of interaction with 3D models, such as camera rotation. This is fundamental to derive a multi-view representation of 3D data useful for aesthetic prediction tasks. Instead, the other platforms often make available just a limited number of 2D renderings, generated and uploaded by the authors of the models. Therefore, those platforms do not enable one to analyse the complete geometry and apperence of a 3D model. Moreover, another fundamental advantage is that the Sketchfab viewer allows one to collect different views of the 3D model under the same rendering parameters, as they are set by 3D artists on the platform itself. Therefore, Sketchfab APIs enable one to perform aesthetic prediction on *static* rendering parameters. This avoids the much relevant bias which would come when changing parameters such as lighting and material across different snapshots of the same 3D model, as it would happen with data from different platforms.

The contribution of this paper is twofold. First, we developed an interactive data visualization platform to navigate the content of Sketchfab. Through the platform, we performed a qualitative and quantitative analysis of Sketchfab numerical scores (views and likes collected by 3D models) and textual information (tags and comments) for different 3D object categories. We used the results of the analysis to select a reasoned subset of annotated models to be included in ViDA 3D, and the metadata to use as supervisory information. Our findings are summarized in Section 3. Besides providing grounds for our design choices while building the dataset, we believe our discussion can also be useful for other applications relying on Sketchfab content. We make the platform publicy available, so that any interested reader can navigate statistical data on Sketchfab and get additional insights.

The second contribution is the introduction of a reasoned dataset which contains the full set of content required for data-driven approaches to 3D aesthetic analysis. In Section 4, we discuss the design choices behind the process of data collection, to identify which models to include; the process of collecting the set of representative 2D images for each 3D model; the metadata to use as supervisory information in learning techniques targeted to aesthetic prediction.

We believe our dataset can be a great resource to support research on the hot and far-from-solved problem of 3D aesthetic prediction.

2. State-of-the-art and beyond

The aesthetics of visual content has long been studied, with contributions from psychology, neuroscience, biology, and cognitive science [KZ04, LBOA04].

2.1. Aesthetic analysis on images

In the field of image analysis, some early research concerned aesthetic prediction on faces and pictorial artworks, especially paintings and photographs. Many works focused on technical rather than aesthetic assessment, as they looked for visual features that mimicked the widely-accepted principles followed by professional photographers, for example rules of composition and color harmony [DJLW06].

Lately, data-driven approaches started replacing rules and handcrafted features, boosted by the growing amount of image data annotated with human preferences. Among them, AVA is a dataset explicitly assembled for large-scale evaluation of attractiveness classification and regression tasks [MMP12]. The images in AVA are annotated with attractiveness scores assigned by users, and accompanied by natural language text. Talebi and Peyman proposed a Convolutional Neural Network (CNN) to predict both technical and aesthetic quality of images, by learning the distribution of human opinion scores on AVA [TM18]. Ma and colleagues defined a Multi-Patch CNN architecture for photo aesthetic assessment, targeted to evaluate both fine-grained details and holistic image layout [MLC17]. Also, recent works on image aesthetic prediction took into account contextual information beyond visual information, including accompanying textual tags and comments, to study the dependence of attractiveness on semantic information [MMP15].

2.2. Perceptual studies on 3D data

For 3D content, early research focused on technical rather than aesthetic assessment, to evaluate for example the fidelity of polygonal objects. Then, shape features such as simmetry and curvature were studied as correlates of beauty and preference [MG14], especially for CAD design and the evolution of aesthetically pleasant geometric objects [BR13]. However, no single measure or combination of measures were found to fully characterize what people considered good and attractive in a 3D object.

Much research about human perception and 3D models dealt with the concept of 3D saliency, as a measure to identify surface regions which attract visual attention [LVJ05, LCSL18]. Perceptual models of 3D content were also investigated in a number of works to decide on good views of a 3D object [SLF*11], with notable applications including the definition of optimal 3D printing directions [ZLP*15]. However, none of these works targeted 3D aesthetic prediction.

Many of the early works about perceptual studies resorted to abstract shapes. In the wake of more recent works [SLF*11], our dataset includes models belonging to categories that are easily recognizable by humans. We expect them to lead to more meaningful results than abstract shapes, since they better characterize the models that are likely to occur in the foreseen applications of 3D aesthetic prediction, such as beautification and content presentation for large datasets.

For the collection of human-generated data, most studies resorted to crowdsourcing platforms such as Amazon Mechanical Turk. The Turkers were assigned tasks such as comparing the goodness of different viewpoints or the saliency of surface points on 3D meshes, then the collected data were filtered for consistency and bias [CSPF09]. Further to recent trends in aesthetic image analysis [MMP15, TM18], we preferred to look for 3D data with human annotations collected in *real* scenarios, as they would better reflect the attitude of users towards 3D models in concrete applications.

Finally, previous works rarely took into account attributes besides geometric shape. Exceptions are the works by Ramaranarayanan et al. on the effect of lighting and materials on visual perception [RFWB07], and by Xiao et al. on the perception of translucent materials [XWG*14]. On the contrary, we deem it important to also focus on appearance properties such as color, texture and material, which play a fundamental role in defining the quality and aesthetics of a 3D model.

2.3. 3D benchmark datasets

In the last twenty years, a number of datasets have been proposed to support 3D shape analysis and machine learning research on 3D data. The earlier datasets were meant to benchmark 3D shape analysis techniques, such as shape retrieval [PSMKF04] and segmentation [CGF09]. A notable initiative is the SHape REtrieval Contest (SHREC)[†], which has been running since 2006, and contributing many datasets to evaluate the effectiveness of algorithms for 3D shape matching, classification, retrieval and related tasks. Lately, the spread of data-driven approaches called for large, annotated datasets. ShapeNet[‡] is an ongoing effort to provide the community with richly-annotated, large-scale datasets of 3D shapes. The ShapeNetCore subset includes about 51,300 single, clean 3D models, covering 55 common objects categories, and with manually-verified category and alignment annotations. ShapeNetSem is a smaller subset, with additional annotations about material composition, and estimated volume and weight. Finally, the PartNet dataset includes fine-grained, hierarchical part annotations on ShapeNet models [MZC*19].

Different photorealistic dataset are available to support scene understanding and applications like autonomous robotic exploration. Among them, Matterport3D contains 10,800 panoramic views from 194,400 RGB-D images of 90 building interior scenes, annotated with the camera poses of each panoramic view, and 2D and 3D semantic segmentations [CDF*17]. Replica is another highly-photorealistic dataset of indoor scenes [SWM*19], used in the embodied AI simulation platform AI Habitat [SKM*19].

Despite the abundance of datasets regarding 3D content, up to our knowledge there is no dataset tailored to support research on 3D aesthetic prediction. The desiderata for such a dataset would include: the availability of a large number of 3D models belonging to human-understandable categories, the presence of metadata useful to assess aesthetic preference, and visual information about both geometric shape and appearance.

2.4. 3D content sharing platforms

Possible sources for getting useful data to design aesthetic prediction pipelines include showcase platforms for 3D content sharing, where artists and developers upload their models. The target applications include gaming, augmented and virtual reality, advertising, animation, movies, and 3D printing. Example platforms are Sketchfab, Quixel Megascans, ArtStation, cgtrader.

Sketchfab[§] is a very popular platform, containing more than 3,000,000 3D models, for publishing and sharing 3D content that nicely responds to the desiderata listed above. The 3D models uploaded by users are organized in human-understandable categories of 3D objects, from cultural heritage objects to pets and animals, from vehicles to military; the models also exhibit a rich and large variation in modeling and rendering styles. The models are endowed with human-generated data, which reflect the attitute of *real* users towards the models presented, including both quantitative scores (the number of views and likes received, along with the date of upload for score normalization) and contextual information such as textual tags and comments. Finally, the models are showcased through an interactive viewer, and the officially realesed APIs enable to rotate the model and grab snapshots from different viewpoints.

Concerning competitor platforms, Quixel Megascans[¶] features a limited number of 3D assets (about 13.750), and does not provide useful information about user preferences, such as likes and views. Therefore, it does not respond to the desiderata for a dataset for aesthetic prediction. ArtStation[‖] includes about 49.800 prod-

[†] https://www.shrec.net/

[‡] https://www.shapenet.org/

[§] http:\www.sketchfab.com

[¶] https://quixel.com/megascans/home

[‖] https://www.artstation.com/

ucts created by 3D artists, and a score based on user ratings. Nevertheless, only a limited number of snapshots and videos are made available, and there is no way to collect additional data about the complete 3D geometry and appearence of the models. cgtrader[**] includes about 1,110,000 3D models in 17 categories. Similarly to the other platforms above, 3D models are showcased through a limited number of 2D media content. Besides, often the same 3D model is presented under very different rendering paramters, which would make it difficult to develop an aesthetic prediction technique taking into account both shape and appearance. On the contrary, Sketchfab interactive viewer and APIs enable the gathering of any number of different views of the 3D model, with rendering parameters preserved across views.

For these reasons, we decided to build our dataset for 3D aesthetic prediction using information collected from the Sketchfab website. To further validate our preference for Sketchfab as a source to collect data, and to decide on the specific 3D models to include in our dataset, we performed a preliminary, in-depth study on the Sketchfab content, presented in the next Section.

3. Dissecting Sketchfab

To design a study about how pleasant the appearance of 3D objects is according to human preferences, the first concern is what 3D models to use for the study. A second concern is what human-generated data to collect for measuring aesthetic preference, as there is no absolute scale for such measurement. To decide on these issues, we developed a platform to visually analyse the Sketchfab content: we analysed the population of different categories, evaluated the number of views and likes received by models, and checked their distribution in each single category. We also quantitatively and qualitative analysed accompanying textual tags and comments. A summary of our findings is reported in this Section. The platform for navigating the complete data is available online[††].

3.1. Data collection

The data for the statistical study was collected in-between March 13 and March 16, 2020. We decided to take into account all the models in Sketchfab having a sufficiently large number of views, because we wanted to select objects observed by a sufficiently large number of people to reduce the bias on preference scores. Therefore, we retrieved metadata for all models with more than 175 views since their upload to the platform. The result was a collection of metadata for about 228K 3D models, with the number of views per model ranging in-between 175 and 16 millions. The data were crawled from the Sketchfab website using its public API, through a java application developed for the purpose.

3.2. Sketchfab content analysis

The 3D models in Sketchfab are labelled as belonging to one of more categories. Figures 1, 2, 3 show some of the models belonging to the three biggest categories, namely *Characters and Creatures*, *Architecture*, and *Cultural Heritage and History*, respectively. The models exhibit substantial intra-class variability. The objects can be modelled or acquired, with our without a background, and with different rendering styles. Also, each category includes a number of different subjects. For example, characters and creatures include realistic humans as well as humanoids; realistic and antropomorphic animals; cartoon models; popular figures; different types of assets such as wings and jackets (Figure 1). Architectural models include, among others, interior design models and buildings, either reproductions of physical ones or featuring in videogames; architectural plans; assets for game and video production (Figure 2). Cultural heritage models feature statues, sculptures, and other works of art; items from real-world capture systems; cultural artefacts, such as a barber chair and a Columbus caravel (Figure 3). The variability makes the task of aesthetic prediction really challenging, and confirms the need for large and curated data.

Distribution of models, views and likes. The table in Figure 4 reports the approximated number and percentage of models belonging to each of the eighteen Sketchfab categories, and, for each category, the number and percentage of views, likes, and comments received from users. The reader can view the exact, non approximated figures, by hovering the mouse on the corresponding plot on the interactive platform. On the platform, the reader can also consult additional statistics about non-subjective geometric quantities (numbers of vertices and faces). At any rate, the metadata on views, likes, and comments are expected to provide more precious supervisory information than the geometric quantities. Indeed, they are explicit, tell-tale signs for the human interaction with and attitude towards the 3D objects.

Figure 5 visually illustrates the quantities in the table. It offers an aggregated view on the size of a category, and the amount of views, likes and comments collected by the models inside that category. One can notice that, in general, the numbers are in proportion, apart for the category *Animals and Pets*. This seems to suggest that, for this category, the human preferences are more dependent on the specific semantic content than for other categories.

The relationship between the number of views and likes for different models is investigated in Figure 6 for the category *Characters and Creatures* (for the other categories and for further exploration of the data, we point the interested reader to the interactive graphs on the project website). Each dot in the scatterplot is a 3D model, whose coordinates are the number of views ($x-$axis) and likes ($y-$axis). A mouse hovering over the dot links to the corresponding Sketchfab model. We observed that outliers in the number of views and/or likes often correspond to animated models and models with sound, for example, the *Mech Drone* model, which also featured on the Sketchfab webpage for some time. This made us decide not to include models with animations in our ViDA 3D dataset, to remove a potential source of bias. In the same figure, we show the placement of two other models, namely *Diorama* and *Mask*, which are not animated, though they received a high number of likes relative to the number of views.

Textual comments. In the AVA dataset, the presence of lengthy comments was reported as a sign of popularity for images. Therefore, we decided to analyse the number and length of comments

[**] https://www.cgtrader.com/

[††] https://www.vitoferrulli.it/sketchanalysis/about.html

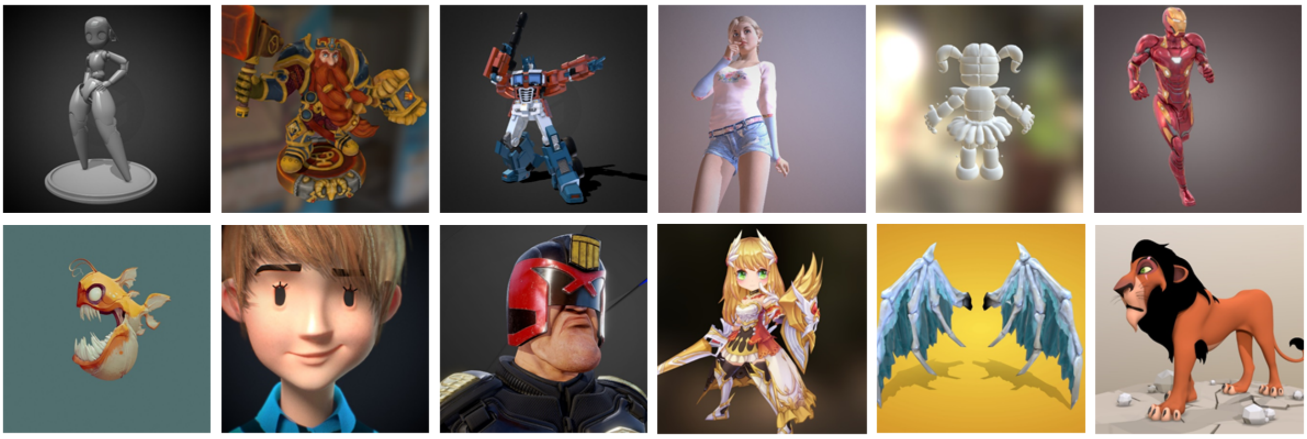


Figure 1: *Examples of 3D models in the "Characters and Creatures" category.*

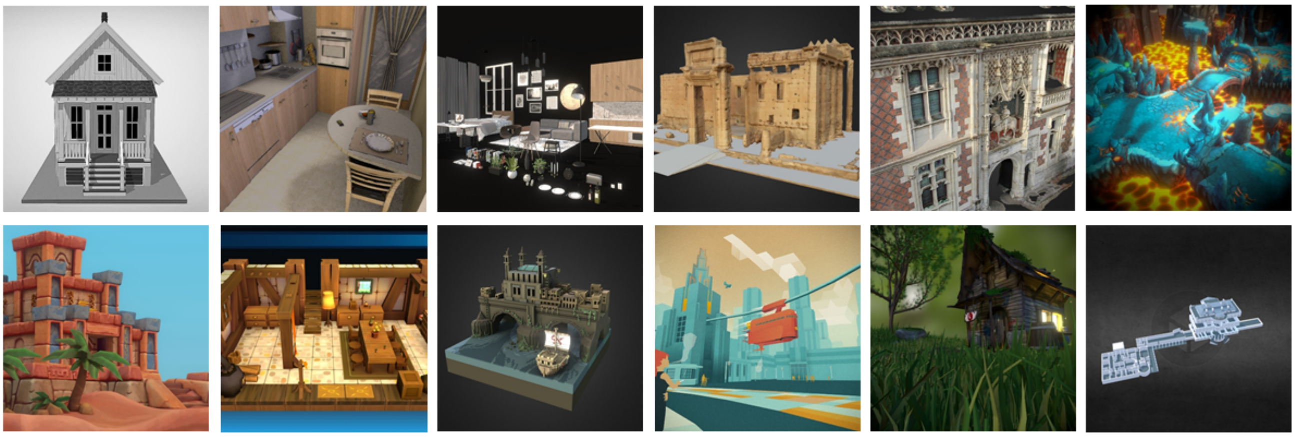


Figure 2: *Examples of 3D models "Architecture" category.*

per 3D model. Though, while comments in AVA are inserted in response to a given task, comments in Sketchfab are freely inserted by users, without any constraint. Therefore, their presence and length is a less reliable parameter. In general, we did not notice any correlation between their number and/or length and the interest that a model drew from users in terms of views and likes.

To go beyond comment length, we mined the text of comments to find the most frequently used words, reported as usual in our website. The results were manually cleaned, to pair synonims, remove stop-words, etc. We believe that the presence of terms related to positive/negative feelings (e.g., *great*) can be leveraged as supervisory information, along with other metadata related to likes. For this reason, we also collected and included comments in our set of metadata, to support multidiciplinary research projects involving both text and visual content analysis.

Textual tags. The textual tags accompanying the 3D models on Sketchfab revealed a potentially precious source of information: they are more reliable than comments, as they are assigned by the authors of the models. To assess whether tags contain terms which can be predictors of human preference, along with the visual content, we analysed the relationship between the presence of particular tags and the number of views and likes. In the scatterplots of Figures 7 and 8, each dot is a tag, whose coordinates are its frequency of occurrence for the models in the category (x-axis) and the number of likes accumulated by the models having that tag (y-axis), for the categories *Characters and Creatures* and *Cultural Heritage and History*, respectively. One can observe that terms which relate to modeling and rendering styles like for example *handpainted* and *stylized* are likely to be attached to models receiving a significant number of likes. Similarly, tags referring to the subject represented (for example, *man*, *girl*, *Paris*, *mithology*) can be predictors of different levels of attractiveness. Similar considerations hold for Figure 9, where the $y-$axis represents the number of views and the category is *Architecture*, and Figure 10, where the coordinates of a tag are given by the numbers of likes and views received by the models having that tag. One can observe that, for example, animals tagged with *lowpoly* are likely to be popular, probably because of their high quality texturing. Also, it is worth observing that the per-

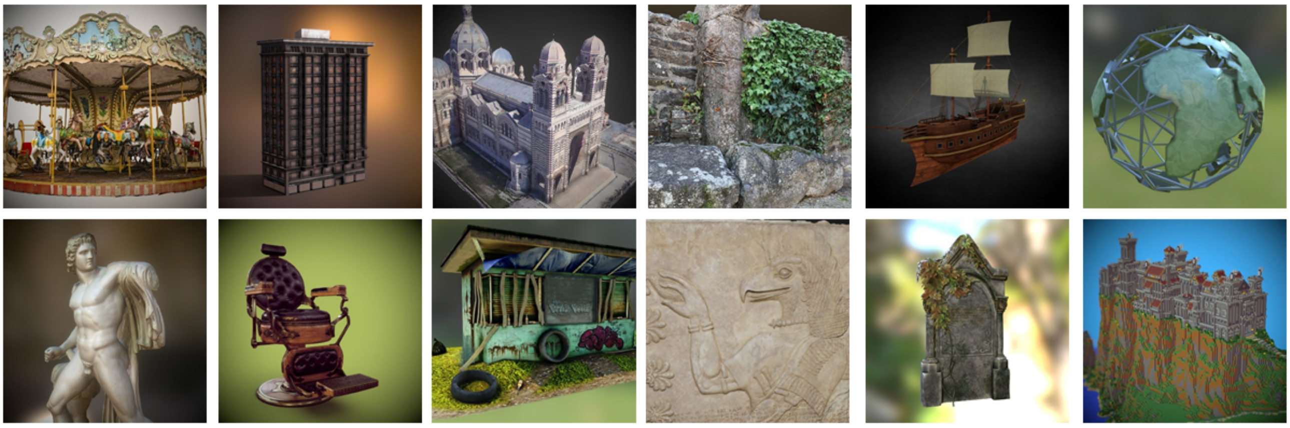
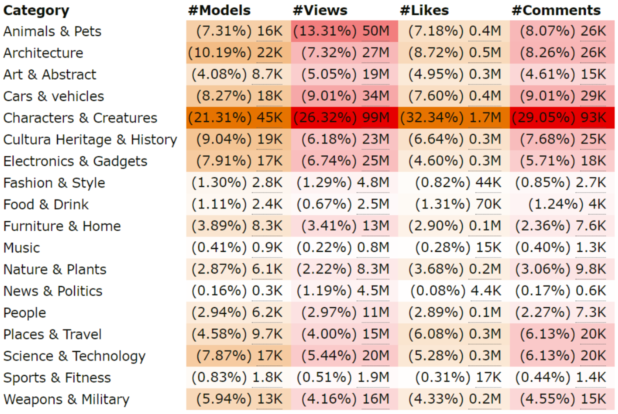


Figure 3: *Examples of 3D models in the "Cultural Heritage and History" category.*



Category	#Models	#Views	#Likes	#Comments
Animals & Pets	(7.31%) 16K	(13.31%) 50M	(7.18%) 0.4M	(8.07%) 26K
Architecture	(10.19%) 22K	(7.32%) 27M	(8.72%) 0.5M	(8.26%) 26K
Art & Abstract	(4.08%) 8.7K	(5.05%) 19M	(4.95%) 0.3M	(4.61%) 15K
Cars & vehicles	(8.27%) 18K	(9.01%) 34M	(7.60%) 0.4M	(9.01%) 29K
Characters & Creatures	(21.31%) 45K	(26.32%) 99M	(32.34%) 1.7M	(29.05%) 93K
Cultura Heritage & History	(9.04%) 19K	(6.18%) 23M	(6.64%) 0.3M	(7.68%) 25K
Electronics & Gadgets	(7.91%) 17K	(6.74%) 25M	(4.60%) 0.3M	(5.71%) 18K
Fashion & Style	(1.30%) 2.8K	(1.29%) 4.8M	(0.82%) 44K	(0.85%) 2.7K
Food & Drink	(1.11%) 2.4K	(0.67%) 2.5M	(1.31%) 70K	(1.24%) 4K
Furniture & Home	(3.89%) 8.3K	(3.41%) 13M	(2.90%) 0.1M	(2.36%) 7.6K
Music	(0.41%) 0.9K	(0.22%) 0.8M	(0.28%) 15K	(0.40%) 1.3K
Nature & Plants	(2.87%) 6.1K	(2.22%) 8.3M	(3.68%) 0.2M	(3.06%) 9.8K
News & Politics	(0.16%) 0.3K	(1.19%) 4.5M	(0.08%) 4.4K	(0.17%) 0.6K
People	(2.94%) 6.2K	(2.97%) 11M	(2.89%) 0.1M	(2.27%) 7.3K
Places & Travel	(4.58%) 9.7K	(4.00%) 15M	(6.08%) 0.3M	(6.13%) 20K
Science & Technology	(7.87%) 17K	(5.44%) 20M	(5.28%) 0.3M	(6.13%) 20K
Sports & Fitness	(0.83%) 1.8K	(0.51%) 1.9M	(0.31%) 17K	(0.44%) 1.4K
Weapons & Military	(5.94%) 13K	(4.16%) 16M	(4.33%) 0.2M	(4.55%) 15K

Figure 4: *Statistics on Sketchfab, related to about* 228*K models downloaded in March 2020.*

formance of some tags is dependent on the model category: for example, *zbrush* is more likely to denote successful models of animals or creatures than architectural objects.

The above examples (and others that can be discovered while playing with our interactive platform) suggest that tags cab be leveraged to define models for aesthetic prediction which couple visual information with additional semantic content. Therefore, tags are included as metadata in our dataset.

4. ViDA 3D

The analysis of the content of Sketchfab presented above helped us design the ViDA 3D dataset, as a resource to support research on 3D aesthetic prediction.

4.1. 3D model selection

Since our aim was to collect models representing a broad range of common shapes, we decided to include models from the three largest categories, namely *Characters and Creatures*, *Architecture*, and *Cultural Heritage and History*, which accounted for about 40% of the data we examined.

To avoid possible biases due to limited showcasing, we only included objects with more than 2000 views and a lifespan between upload and retrieval longer than 3 months. This led to collecting data about around 7,700 3D models. In line with the findings in Section 3, we also removed animated models from the data.

4.2. Visual content

There is no canonical representation for visual 3D content. While volumetric representation are often cumbersome to manage, and coloured point clouds can be inadequate to appreciate 3D models visually, sets of 2D views are a poweful representation for 3D data. View-based representations are common to many 3D learning techniques, for example for 3D shape classification [RROG18] and semantic segmentation [DN18]. For these applications, 2D views were capable of conveying relevant information about both 3D geometry and appearance. Therefore, to support research on content-based 3D aesthetic prediction, for each 3D model we included in the dataset 46 high quality 2D renderings, corresponding to snapshots taken under different camera positions.

Figure 11 illustrates the process of views collection, for a subset of camera positions. Starting from a seed position (the viewpoint set by the model author), we define the rotations of the camera by 10°, 45°, and 90° along the horizontal and vertical axes as a generating set of rigid transformations. We pick 46 rotations among the composition of transformations in the generating set, to produce the 2D renderings of the 3D model. The views are selected in order to investigate accurately around the viewpoint chosen by the author, and its antipodal point. The final set of snapshots is depicted for an example model in Figure 12. We get 46 renderings for each model, up to a total number of images in the dataset greater than $350,000$. In the current versione of the dataset, the image resolution is 1024×1024.

We used as the seed position the one chosen by the author while

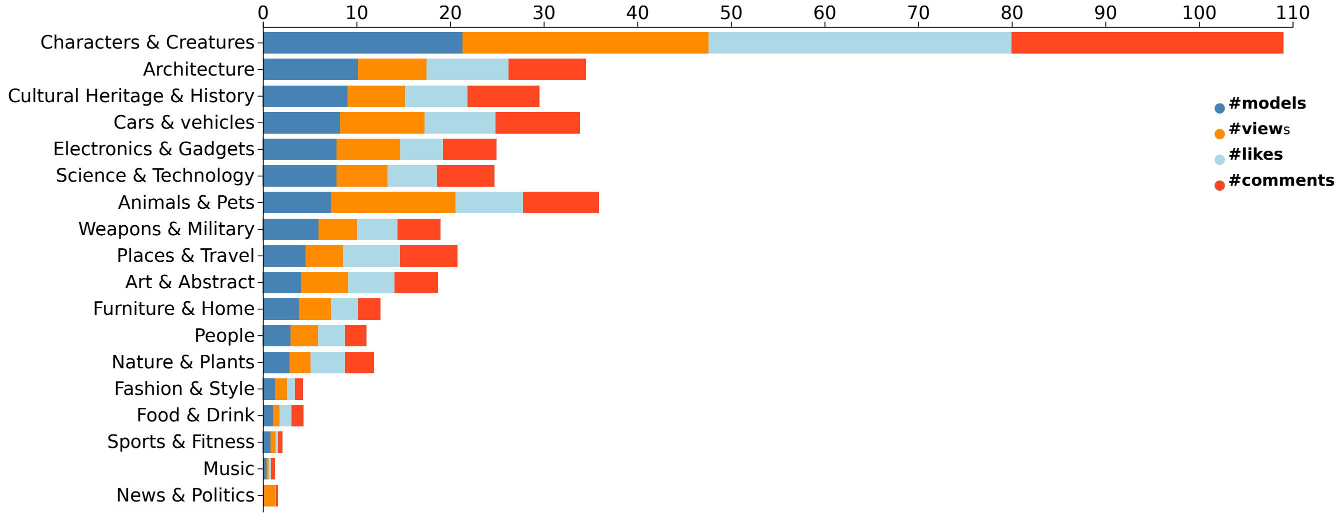


Figure 5: *Statistics on Sketchfab, related to about 228K models downloaded in March 2020: visual comparison on the distribution of models, likes, views and comments in different categories.*

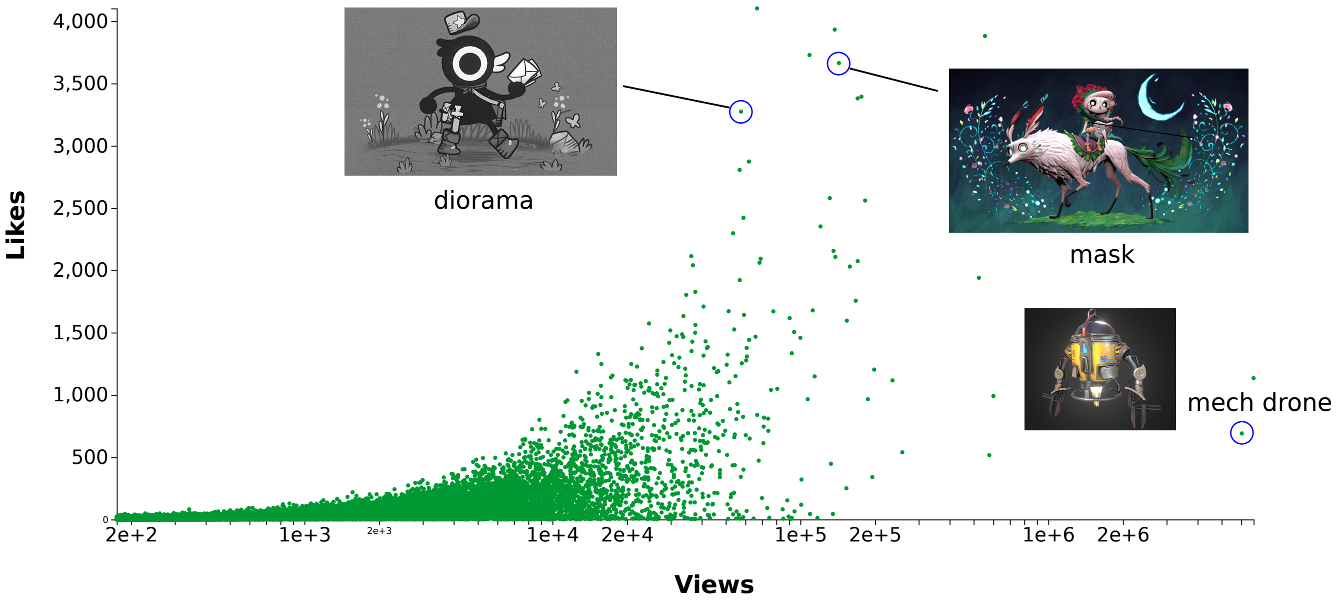


Figure 6: *The relationship between the number of views and likes for different models in the category Characters and Creatures. Each dot in the scatterplot is a 3D model.*

uploading the model to Sketchfab, because it is likely the *best* one to appreciate the model visually, as the author is supposed to showcase an initial view that valorizes his/her creation. The rotations by 10° are meant to augment the visual information carried by the author-defined position. This is also consistent with a study on preference prediction on face images, suggesting that the first photo uploaded by users is the one that mostly affects the like/dislike reaction [RTVG16]. The rotations by higher degrees (45° and 90°) are meant to capture different perspectives on the 3D model.

The set of 46 renderings are expected to provide a good summary of the visual content of the corresponding 3D model, concerning both shape and appearence. Indeed, the renderings provide additional information about the complete 3D geometry of the model, which would not be revealed by a limited number of snapshots. Also, as the rendering parameters remain fixed across all views, machine learning and deep learning techniques trained on such data would learn representations which correctly take into account both geometry and appearance.

Of course, some of the views may show little significance depending on the model at hand, either for technical or semantic rea-

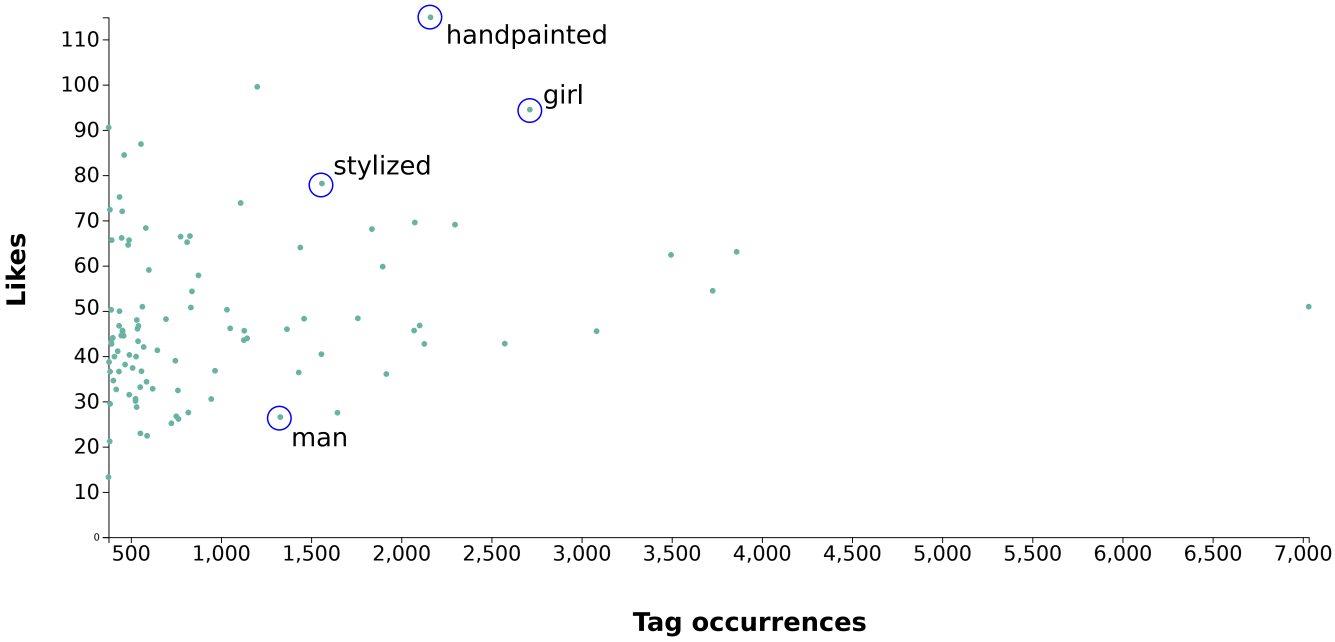


Figure 7: *The relationship between the occurrence of tags and the number of likes in the category "Characters and Creatures". Each dot in the scatterplot represents a tag.*

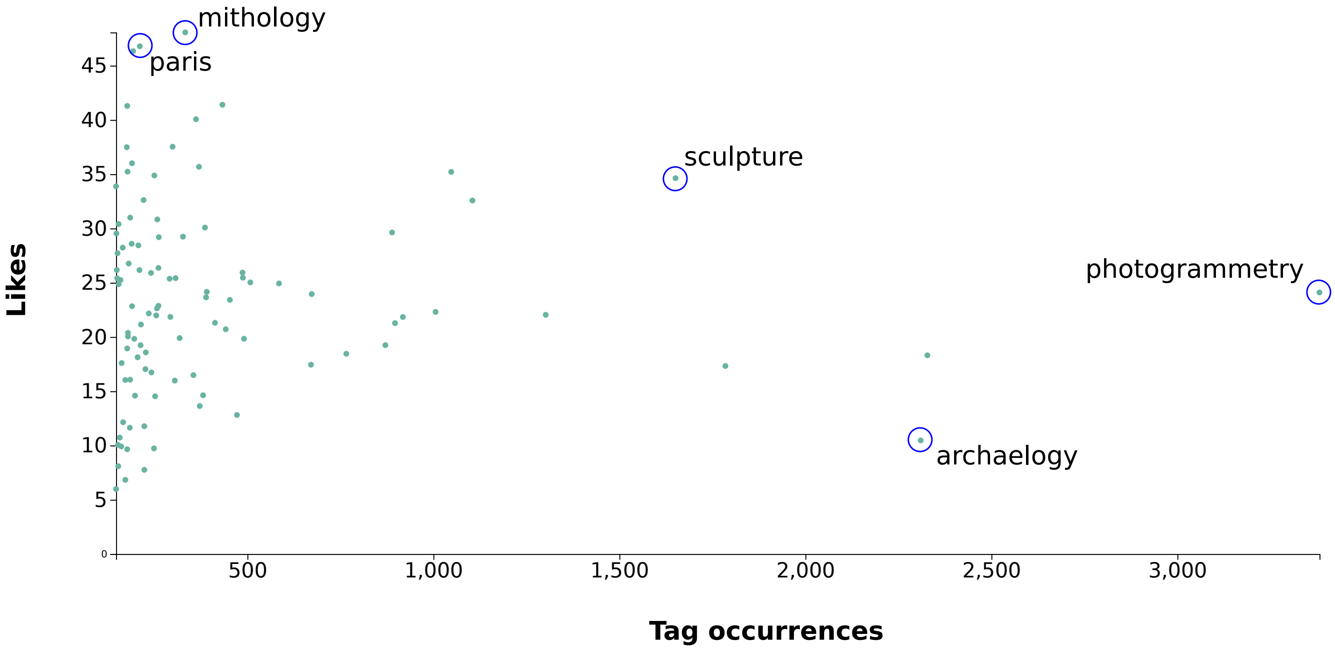


Figure 8: *The relationship between the occurrence of tags and the number of likes in the category "Cultural Heritage and History". Each dot in the scatterplot represents a tag.*

sons. A selection of the most meaningful renderings is expected to be part of the pipeline of 3D aesthetic prediction algorithms, either as a preprocessing step, or as a part of the learning procedure.

4.3. Aesthetics-related metadata

In the light of preference prediction studies on other types of media [GUB*15, MMP15], we associate with each 3D model two different scores as supervisory information to train and evaluate

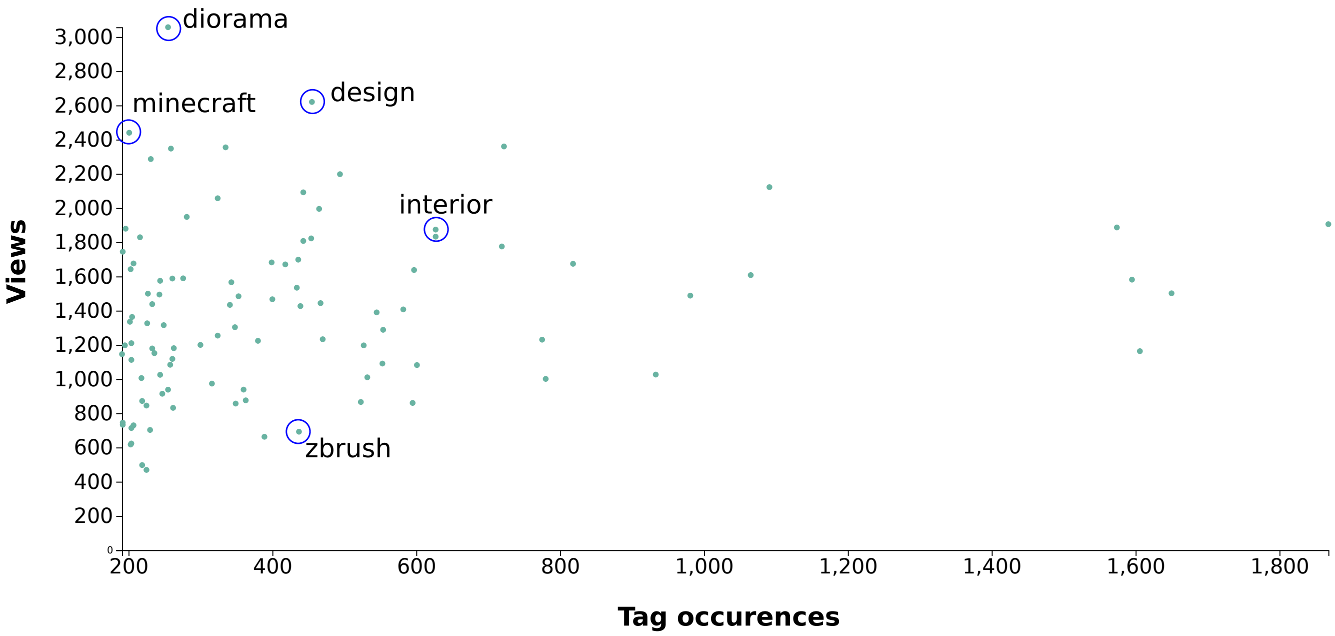


Figure 9: *The relationship between the occurrence of tags and the number of views in the category "Architecture". Each dot in the scatterplot represents a tag.*

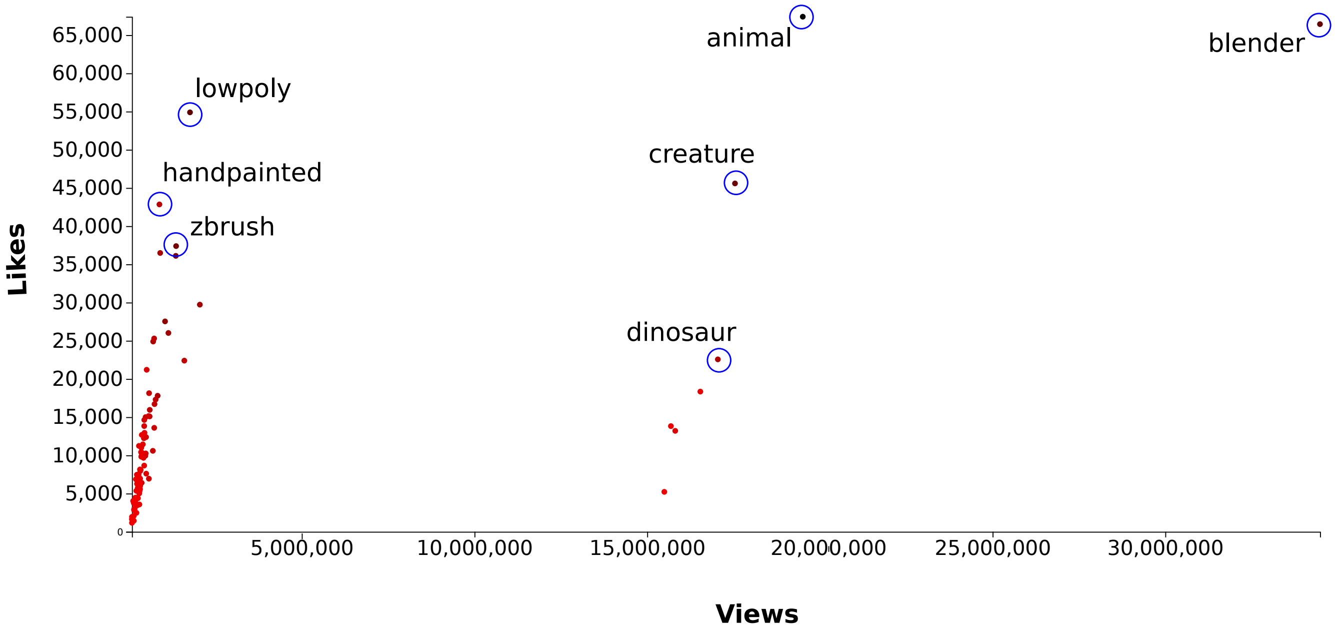


Figure 10: *The relationship between the the number of likes and views for tags in the category "Animals and Pets". Each dot in the scatterplot represents a tag.*

learning-based 3D aesthetic prediction methods: the average number of likes per time span, and the ratio between the number of likes and the number of views.

Nevertheless, there is no consensus in the literature as to which measures to use for aesthetic prediction. The measures proposed for image aesthetic prediction and social media popularity also include, besides ratios of quantities such as the numbers of views and likes, rankings based on pairwise comparisons [MVL*19]. Other works also proposed the use of contextual information included in textual comments to improve on the interpretability of results. As observed in Section 3, textual tags are possible predictors of human

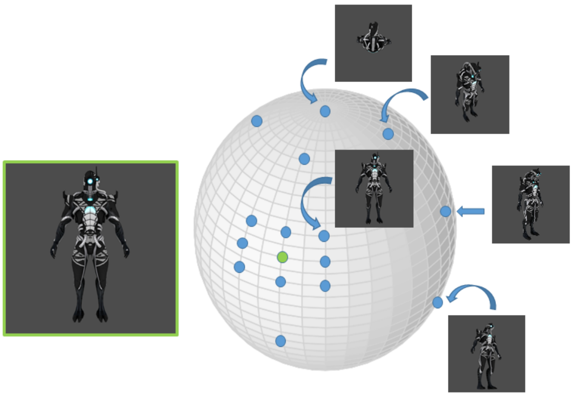


Figure 11: *The process of generating different renderings for 3D models, according to different camera rotations.*

preference, and terms mined from comments are correlated with user feelings towards the models. Therefore, textual data can be leveraged to define techniques for aesthetic prediction which couple visual information with semantic content.

Therefore, to support the definition of additional measures as supervisory information, and to support multi-disciplinary research on 3D aesthetic prediction, for each 3D model the metadata in our dataset include the individual quantitive parameters plus the complete available textual data: the lifespan of the model between upload and retrieval (for possible normalizations); the number of times the model has been viewed; the number of likes received; the textual tags that accompany the model; and the comments by Sketchfab users.

4.4. Implementation details

The implementation of the data collection is based on Javascript libraries. A client-server architecture generates the 2D renderings of the 3D model. The client uses the official Sketchfab viewer library to grab the snapshots, and an Express server stores them in order to ignore the browser limitation on the local file system. The metadata are downloaded via the Sketchfab web API system and organized into a single JSON file.

5. Conclusions

We presented the ongoing effortAlso, the paper now better underlines that, as observed by Reviewer 3, our contribution also includes the development of a publicly available visualization platform, which can be used to support other applications relying on Sketchfab data. to build the first 3D dataset for aesthetic analysis of 3D models, including both visual data and metadata. The data are collected from Sketchfab, a popular platform for sharing of 3D models. We designed the dataset after an in-depth statistical analysis of the Sketchfab content, whose main findings we illustrated in this paper. Since the figures can be a useful source of information for researchers interested on using data from Sketchfab, we made our data visualizations publicly available through an interactive website.

Since the visual appearance plays a fundamental role in aesthetic analysis, for each 3D model our dataset includes 46 high-quality renderings, which convey information about the model appearance from different viewpoints. The metadata to support supervised learning strategies include number of views and number of likes, which are quantities useful to assess how popular and how pleasant the model was for the users. Additional metadata include tags and comments, which can be leveraged by multi-disciplinary approaches to 3D aesthetic prediction.

The current version of the dataset is made up of more than 350.000 2D renderings and associated metadata, about more than 7.000 3D models belonging to 3 different categories. The first public release of our dataset is foreseen before the end of November.

While similar datasets are available for images, to our knowledge this is the first attempt to create a benchmark for aestethic prediction on 3D models. Our ambition is to contribute the Computer Vision and Computer Graphics communities with a valuable resource for data-driven research on the emerging field of aestethic analysis, a topic partially explored for images, but still poorly investigated for 3D models.

References

[BR13] Bergen S., Ross B.: Aesthetic 3D model evolution. *Genet Program Evolvable Mac 14* (2013), 339–367. 2

[CDF*17] Chang A., Dai A., Funkhouser T., Halber M., Niessner M., Savva M., Song S., Zeng A., Zhang Y.: Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)* (2017). 3

[CGF09] Chen X., Golovinskiy A., Funkhouser T.: A benchmark for 3D mesh segmentation. *ACM Transactions on Graphics (Proc. SIGGRAPH) 28*, 3 (Aug. 2009). 3

[CSPF09] Chen X., Saparov A., Pang B., Funkhouser T.: Schelling points on 3D surface meshes. *ACM Trans. Graph.* (2009). 3

[DJLW06] Datta R., Joshi D., Li J., Wang J.: Studying aesthetics in photographic images using a computational approach. In *Proc. ECCV, Lecture Notes in Computer Science* (2006), vol. 3953, Wiley Online Library. 2

[DN18] Dai A., Niessner M.: 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *ECCV 2018: 15th European Conference on Computer Vision* (09 2018), pp. 458–474. 2, 6

[GUB*15] Gelli F., Urricchio T., Bertini M., Del Bimbo A., Chang S.-F.: Image popularity prediction in social media using sentiment and context features. In *ACM International Confence on Multimedia* (2015), pp. 907–910. 1, 8

[KZ04] Kawabata H., Zeki S.: Neural correlates of beauty. *Neurophysiology 91* (2004), 1699–1705. 2

[LBOA04] Leder H., Belke B., Oeberst A., Augustin D.: A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology 95*, 4 (2004), 489–508. 2

[LCSL18] Lavoué G., Cordier F., Seo H., Larabi M.-C.: Visual attention for rendered 3D shapes. 191–203. 3

[LVJ05] Lee C. H., Varshney A., Jacobs D. W.: Mesh saliency. In *ACM SIGGRAPH 2005 Papers*. 2005, pp. 659–666. 3

[MG14] Miura K., Gobithaasan R.: Aesthetic curves and surfaces in Computer Aided Geometric Design. *International Journal on Automation Technology 8*, 3 (2014), 304–316. 2

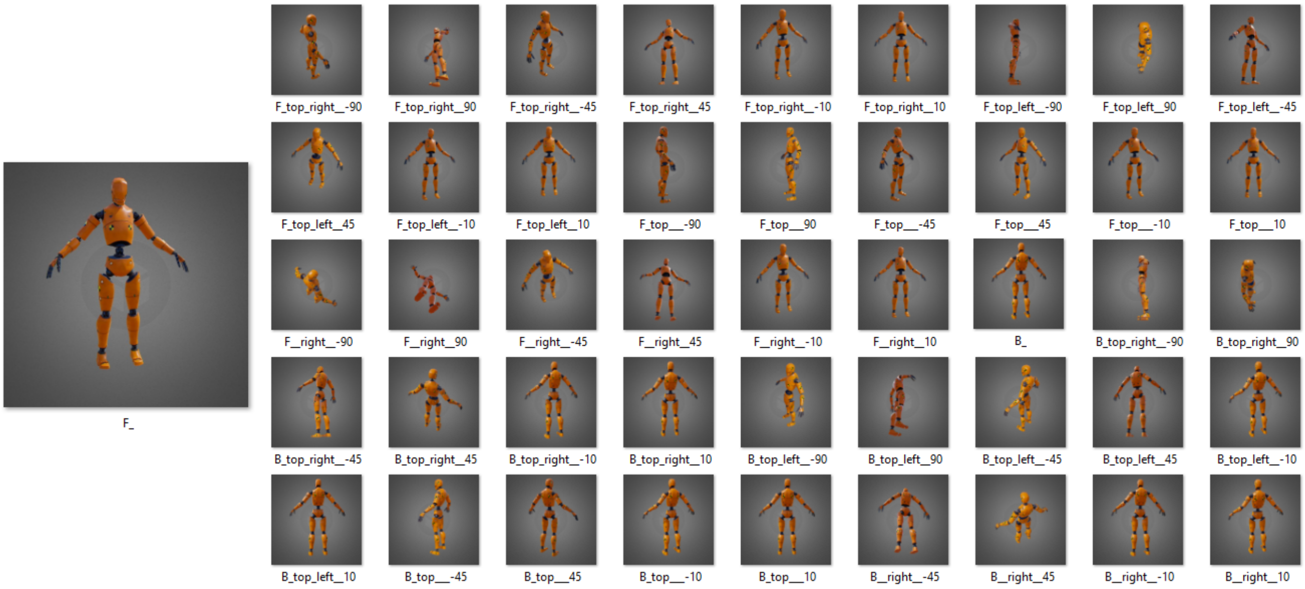


Figure 12: *The rendering from the model position selected by the model author (left), and the additional 45 renderings in the ViDA 3D dataset.*

[MLC17] Ma S., Liu J., Chen C.: A-Lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 722–731. 2

[MMP12] Murray N., Marchesotti L., , Perronnin F.: AVA: A large-scale database for aesthetic visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition* (2012), pp. 2408–2415. 1, 2

[MMP15] Marchesotti L., Murray N., Perronnin F.: Discovering beautiful and ugly attributes for aesthetic image analysis. *International Journal of Computer Vision 113* (2015), 246–266. 2, 3, 8

[MVL*19] Ma N., Volkov A., Livshits A., Pietrusinski P., Hu H., Bolin M.: An universal image attractiveness ranking framework. In *IEEE Winter Conference on Applications of Computer Vision* (2019). 8

[MZC*19] Mo K., Zhu S., Chang A. X., Yi L., Tripathi S., Guibas L. J., Su H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019). 3

[PSMKF04] Philip Shilane P., Min P., Kazhdan M., Funkhouser T.: The Princeton Shape Benchmark. In *Proc. Shape Modeling International* (2004). 3

[RFWB07] Ramanarayanan G., Ferwerda J., Walter B., Bala K.: Visual equivalence: towards a new standard for image fidelity. *ACM Transactions on Graphics (TOG) 26*, 3 (2007), 76–es. 3

[RROG18] Roveri R., Rahmann L., Oztireli A., Gross M.: A network architecture for point cloud classification via automatic depth images generation. In *CVPR 2018: IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 4176–4184. 2, 6

[RTVG16] Rothe R., Timofte R., Van Gool L.: Some like it hot - Visual guidance for preference prediction. In *Proc. Computer Vision and Pattern Recognition (CVPR)* (2016). 7

[SKM*19] Savva M., Kadian A., Maksymets O., Zhao Y., Wijmans E., Jain B., Straub J., Liu J., Koltun V., Malik J., Parikh D., Batra D.: Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2019). 3

[SLF*11] Secord A., Lu J., Finkelstein A., Singh M., Nealen A.: Perceptual models of viewpoint preference. *ACM Transactions on Graphics 109* (2011). 3

[SWM*19] Straub J., Whelan T., Ma L., Chen Y., Wijmans E., Green S., Engel J. J., Mur-Artal R., Ren C., Verma S., et al.: The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797* (2019). 3

[TM18] Talebi H., Milanfar P.: NIMA: Neural image assessment. *IEEE Transactions on Image Processing 27*, 8 (2018), 3998–4011. 2, 3

[XWG*14] Xiao B., Walter B., Gkioulekas I., Zickler T., Adelson E., Bala K.: Looking against the light: How perception of translucency depends on lighting direction. *Journal of vision 14*, 3 (2014), 17–17. 3

[ZLP*15] Zhang X., Le X., Panopoulou A., Whiting E., Wang C. C. L.: Perceptual models of preference in 3D printing direction. *ACM Trans. Graph. 34*, 6 (2015), 215:1–215:12. 3