— Supplementary Material — Textured Mesh Quality Assessment: Large-Scale Dataset and Deep Learning-based Quality Metric

YANA NEHMÉ JOHANNA DELANOY FLORENT DUPONT JEAN-PHILIPPE FARRUGIA PATRICK LE CALLET GUILLAUME LAVOUÉ

This supplementary material is organized as follows. In section 1, we provide visual examples of distorted stimuli from our dataset of textured meshes, along with their distortion parameters. Section 2 describes our large-scale subjective experiment in crowdsourcing as well as the pilot study we relied on. Section 3 provides the parameters of the image quality metrics that we compared to our proposed metric. Section 4 shows the results of our mesh characterization measure applied on several viewpoints. Section 5 provides additional results of Graphics-LPIPS when using different pooling strategies and results on each individual fold. Finally, Section 6 shows the distribution of the predicted quality scores of all the stimuli of our dataset, and provides additional analysis on the impact of distortion interactions and content characteristics on the perceived quality of textured meshes, along with the complete ANOVA table.

1 DATASET GENERATION

We produced a large-scale textured meshes quality assessment dataset composed of over 343k distorted meshes derived from 55 source models each associated with 6250 distorted versions generated from combinations of 5 real-world compression-based distortions applied with different strengths.

1.1 Source preparation

Our source models were collected from sketchFab, an open source online repository for publishing and sharing 3D content. For some models, we had to modify the object files to restore the correct material library files and texture images. For models that were non-manifold and contained zero-area triangles, we fixed this manually using meshLab (https://www.meshlab.net), thus ensuring that all models in the dataset have the same properties.

A few models had multiple texture images. We manually baked these images into a single texture (JPEG image of size 2048x2048) using Blender. We made sure that we got the same visual rendering. This operation facilitates the application of the texture distortions (texture compression and sub-sampling) in the following (distortions applied on 1 image instead of several). Thus all the models in the dataset are represented similarly: by an OBJ file, a material file and a texture image (JPEG image of size 2048x2048).

1.2 Distorsions

The distortions represent (1) the level of detail simplification applied with 10 strengths obtained by uniformly reducing the number of mesh faces ($LoD_{simpL} \in [L1, L10]$, where $L10$ is the most degraded level), (2) the model position quantization ($qp \in [7, 11]$), (3) the texture coordinates quantization ($qt \in [6, 10]$), (4) the texture sub-sampling ($T_S \in \{512\times512, 712\times712, 1024\times1024, 1440\times1440, 2048\times2048\}$), and (5) the texture compression ($T_Q \in \{10, 25, 50, 75, 90\}$). Figures 1 to 10 show visual examples of the generated distorted stimuli along with their distortion parameters.

Author's address: Yana Nehmé; Johanna Delanoy; Florent Dupont; Jean-Philippe Farrugia; Patrick Le Callet; Guillaume Lavoué.

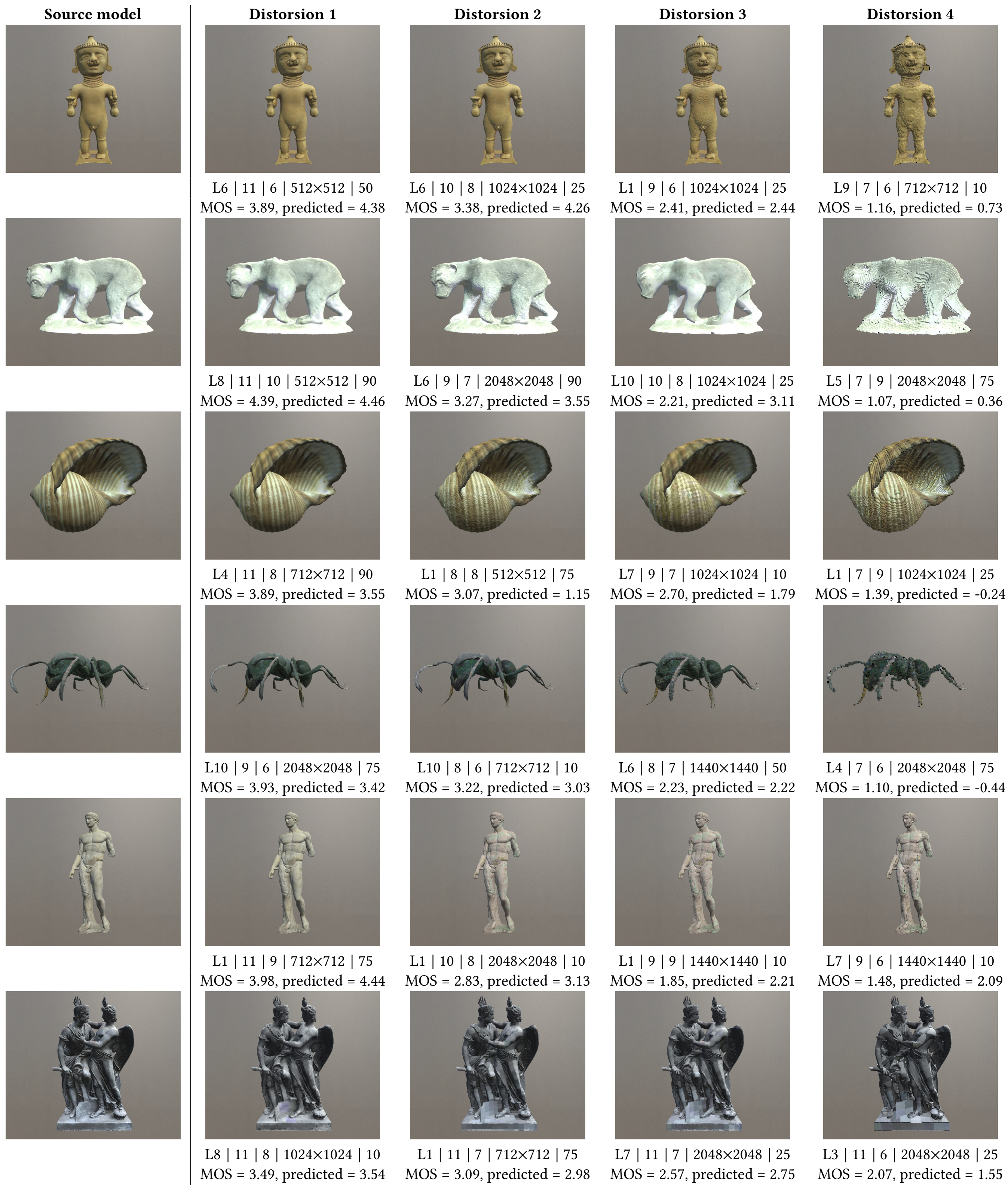


Fig. 1. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

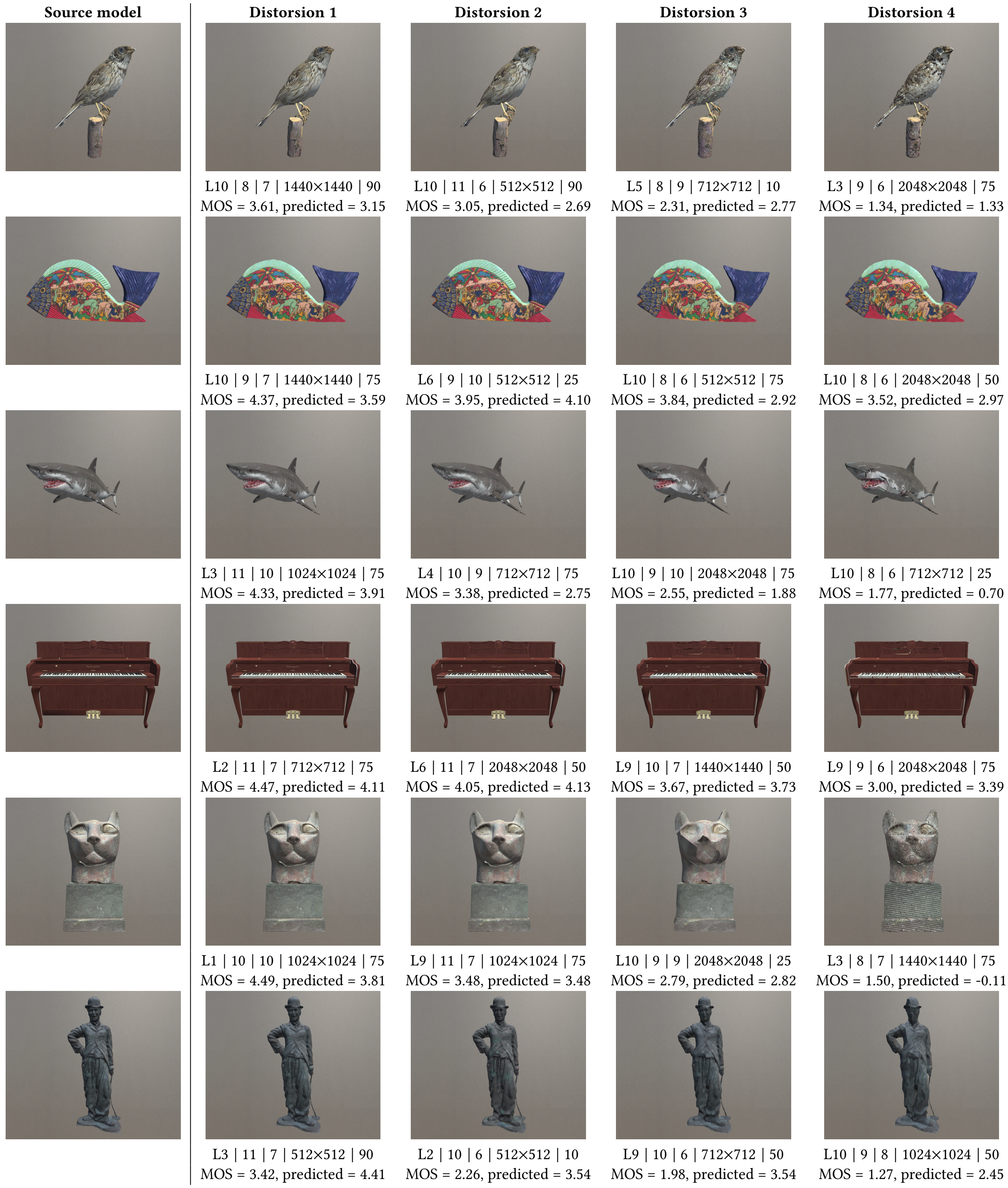


Fig. 2. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

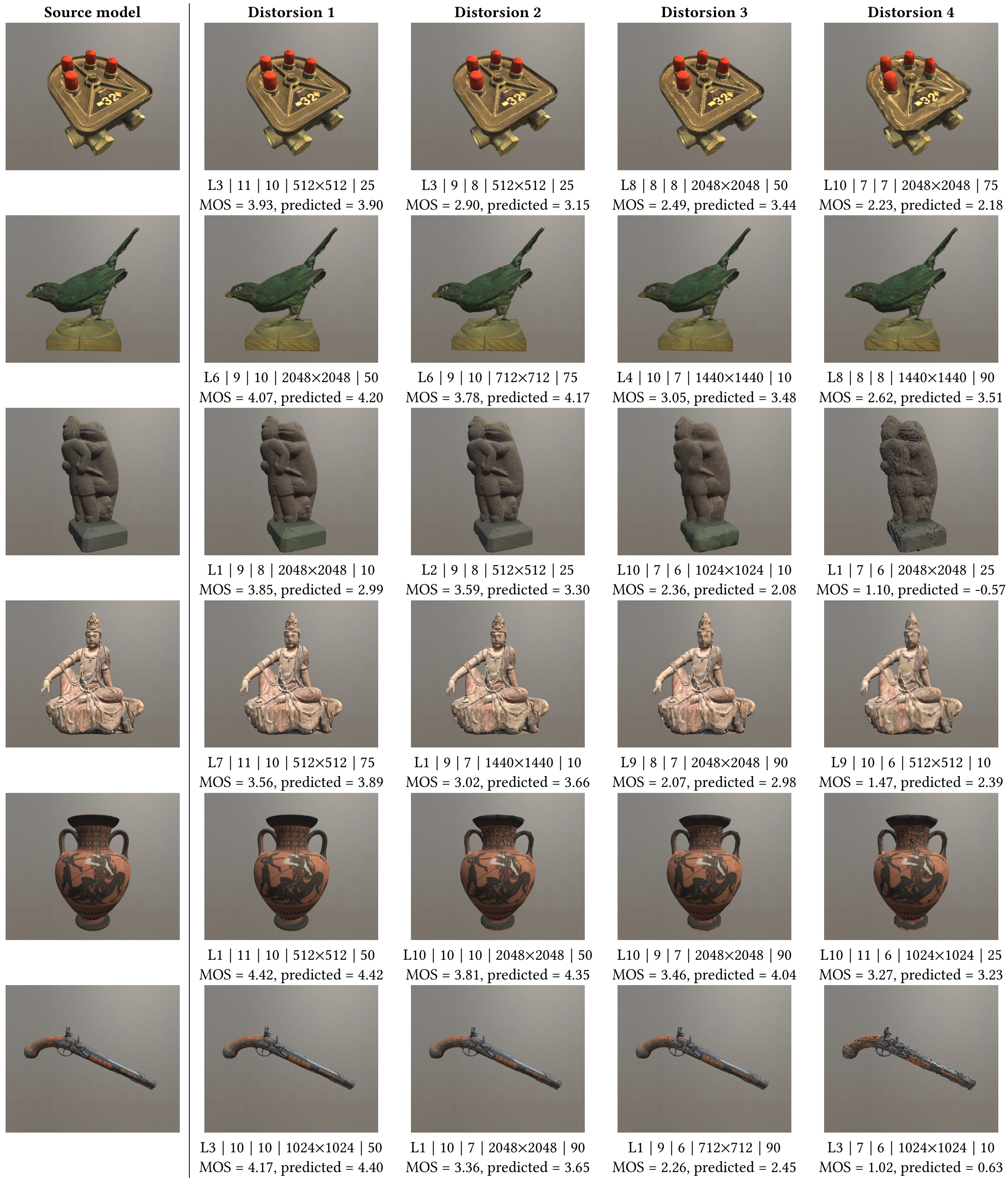


Fig. 3. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

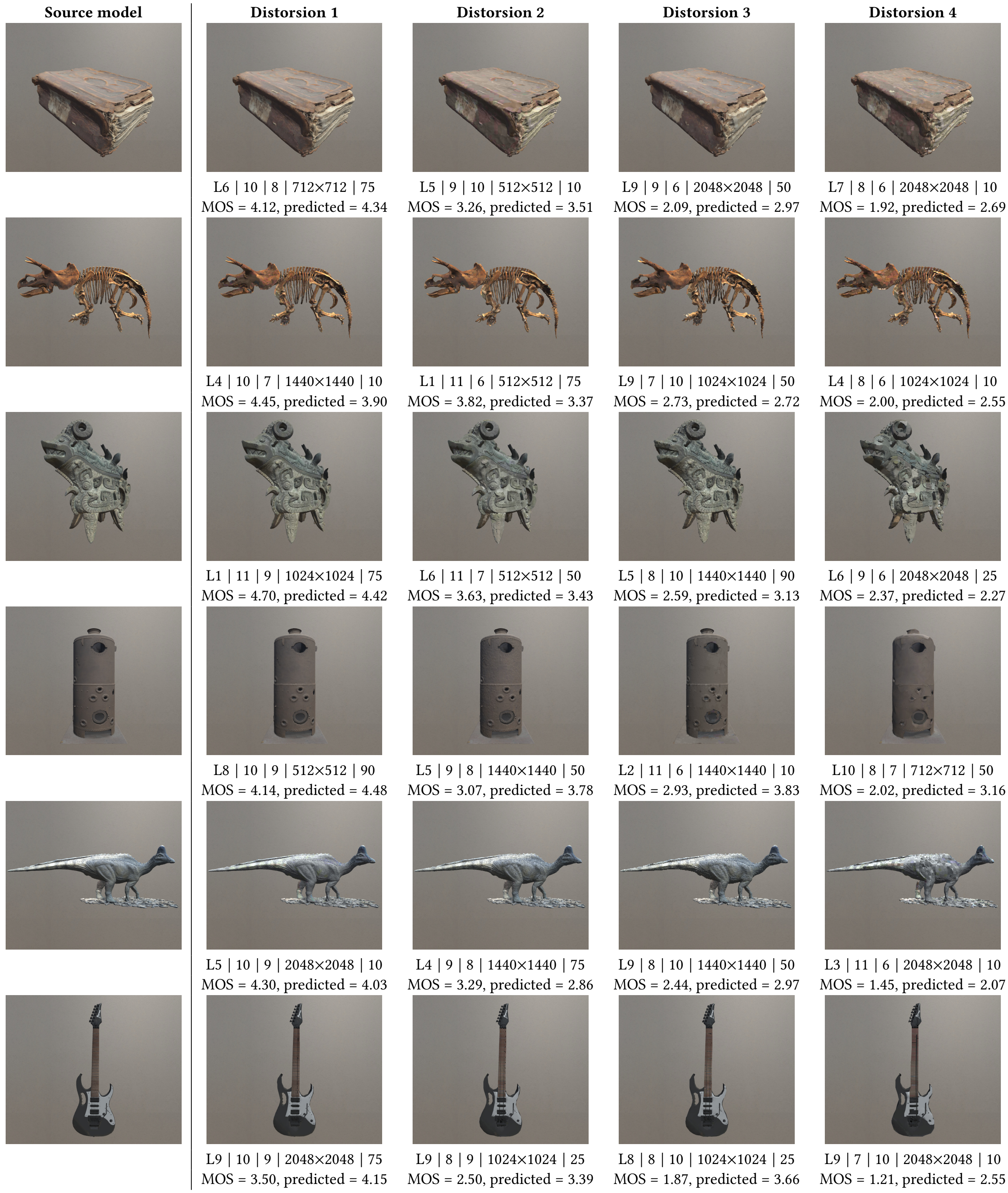


Fig. 4. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

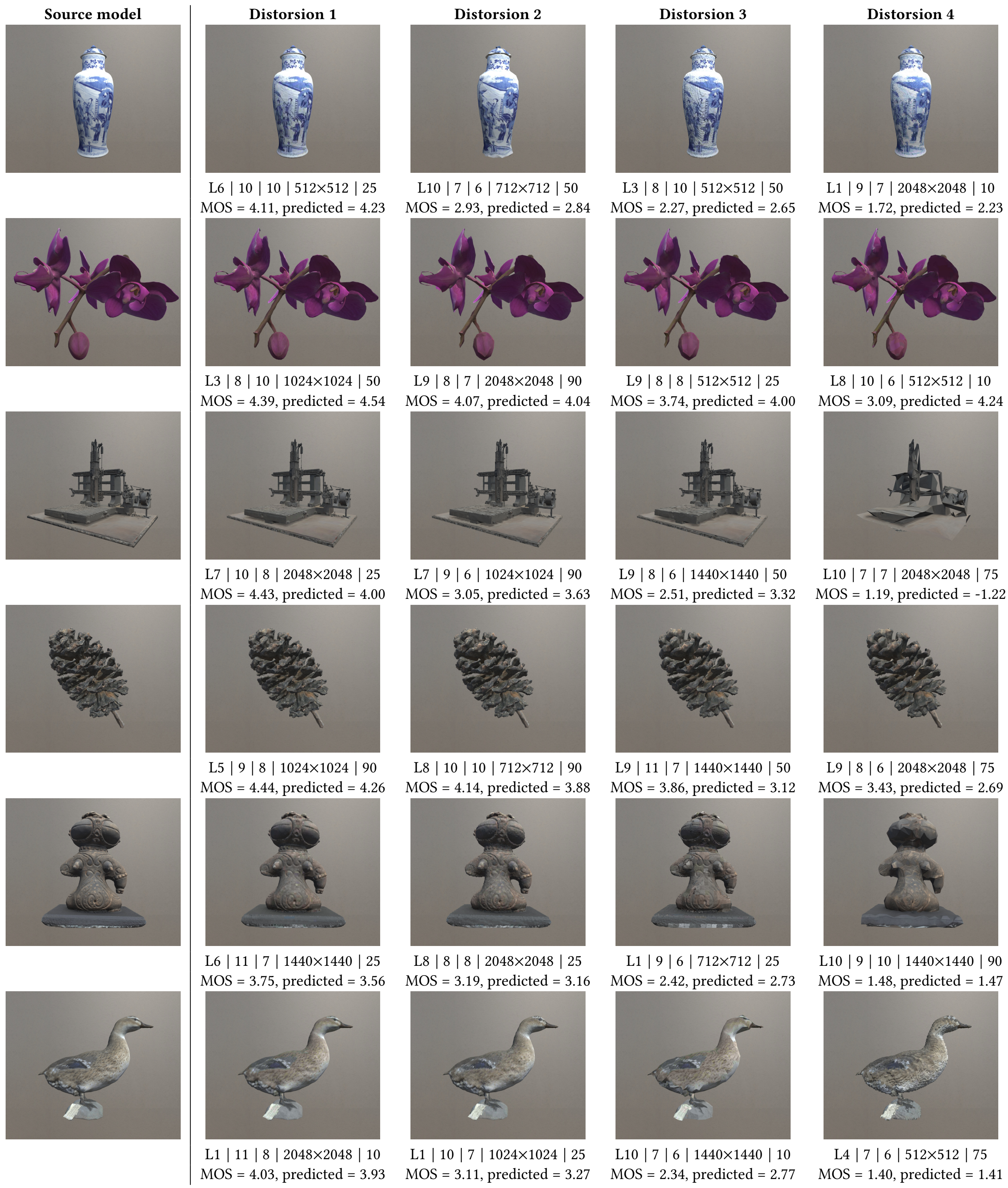


Fig. 5. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

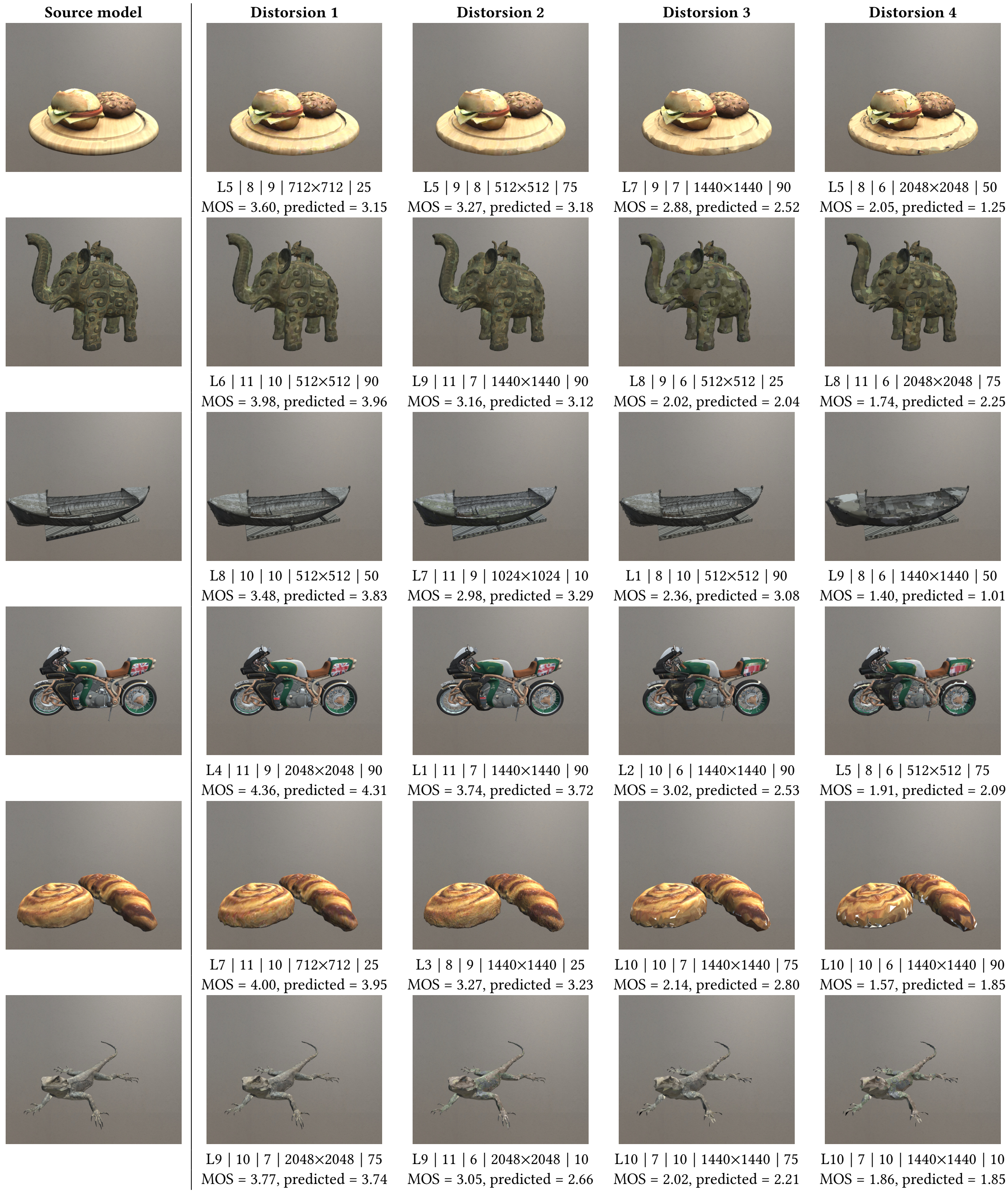


Fig. 6. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

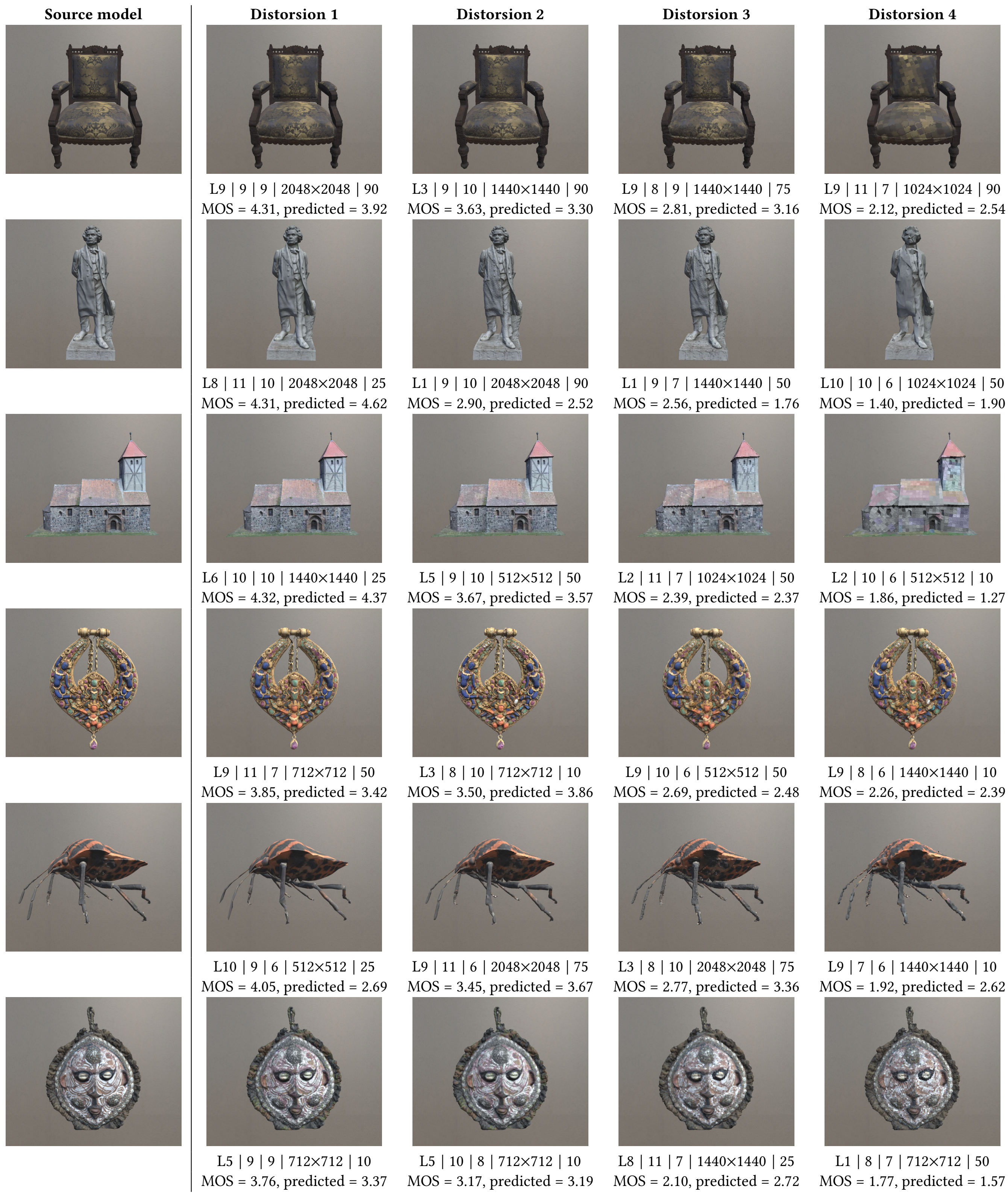


Fig. 7. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

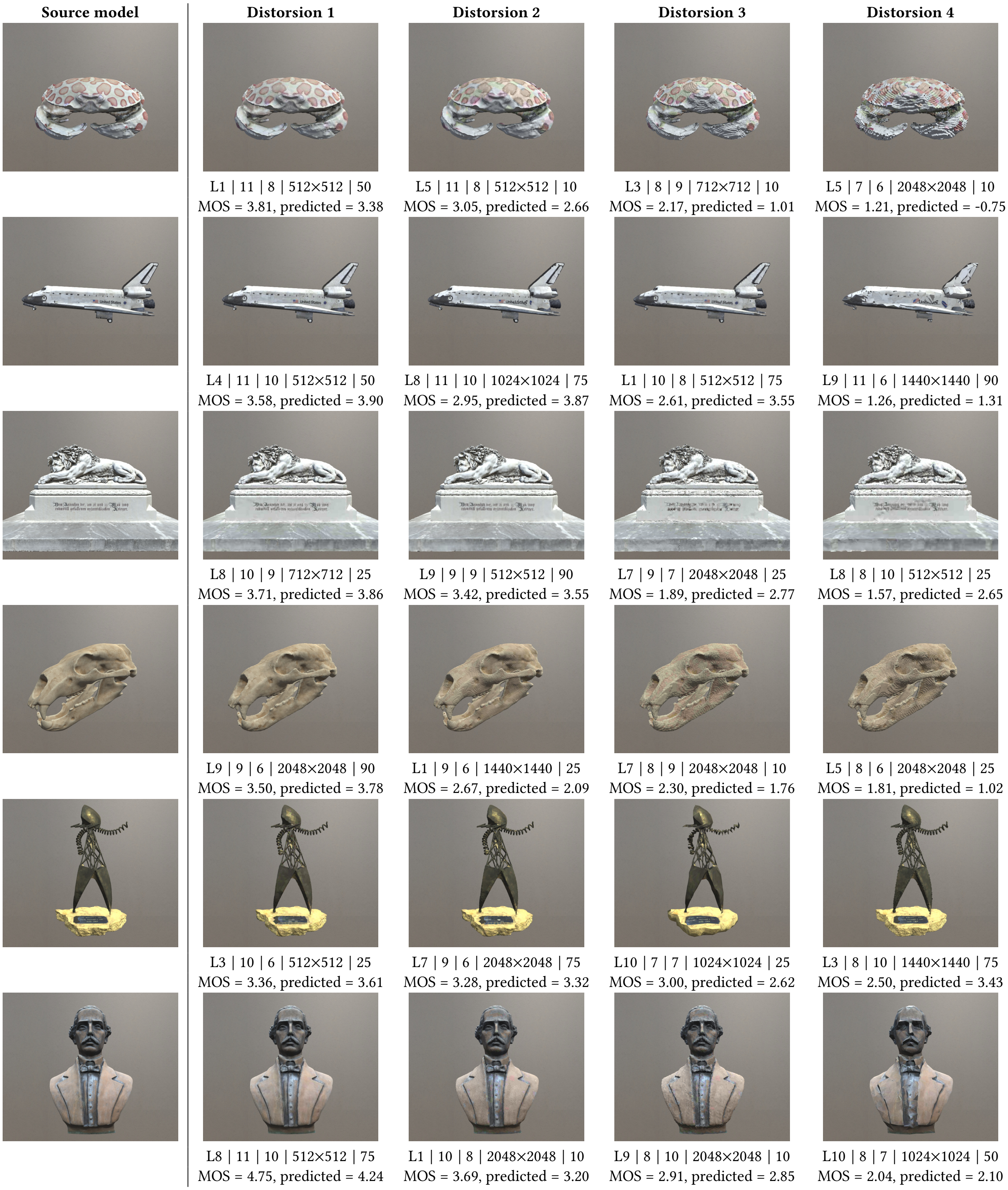


Fig. 8. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

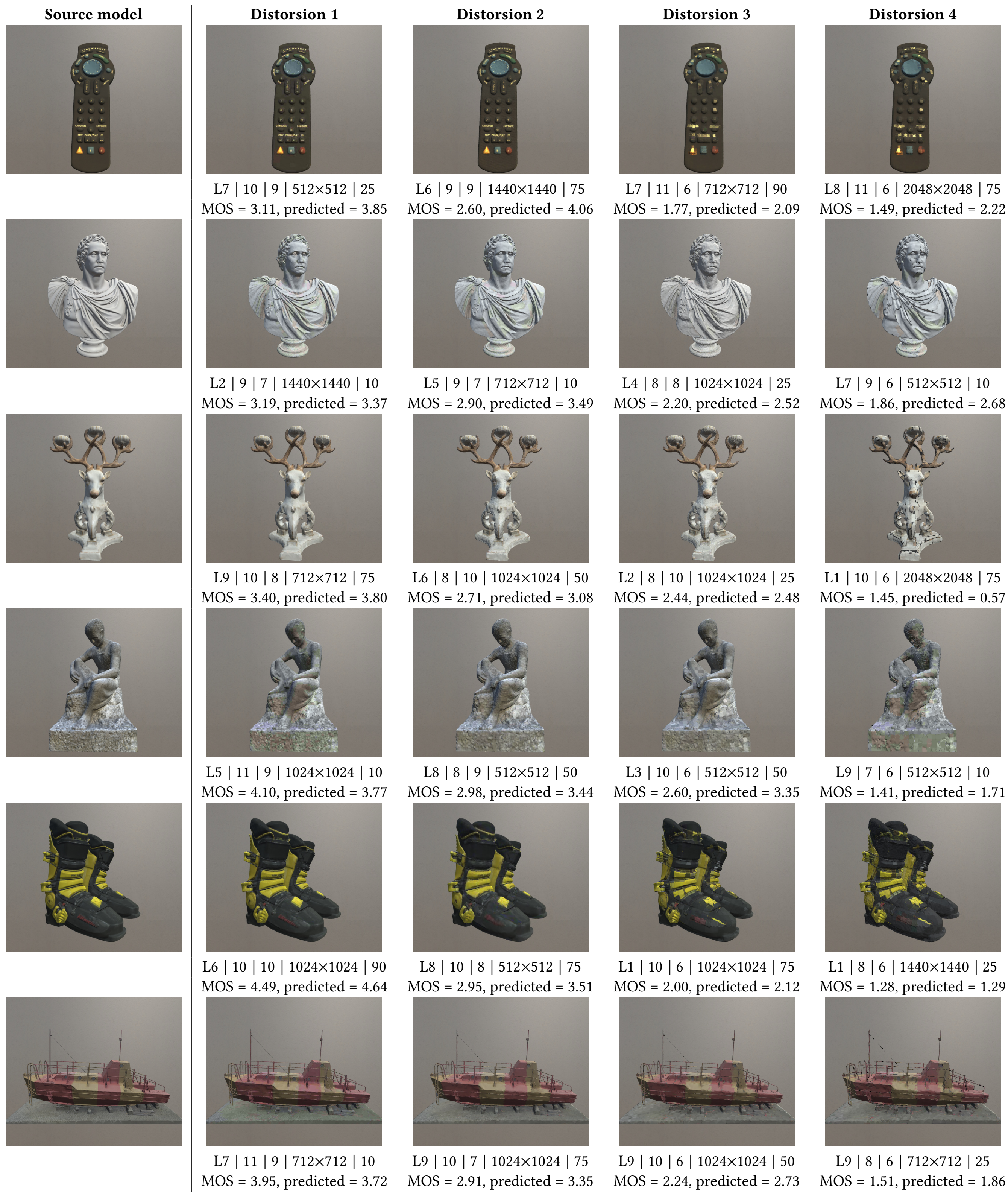


Fig. 9. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

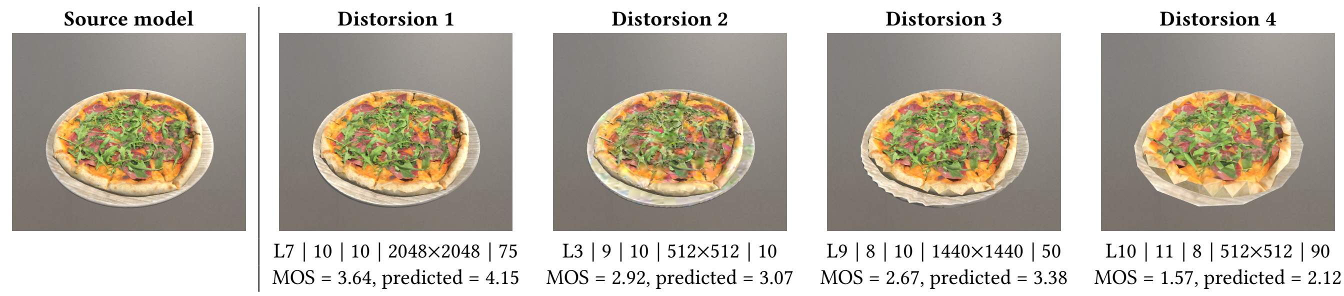


Fig. 10. Examples of stimuli: left-most column is the reference object, the remaining images are randomly sampled distorsions, from the least annoying one (according to MOS) up to the most annoying one. Acronyms refer to LoD_{simpL} | qp | qt | T_S | T_Q

2 SUBJECTIVE EXPERIMENT

We developed our own web platform to conduct the large-scale subjective experiment in crowdsourcing, based on the DSIS method. The crowdsourcing service was used only to recruit participants using Prolific[1], an internet marketplace that provides tens of thousands of trusted participants. We illustrate in the following the successive stages/steps of our experiment. To run the experiment, only a web browser with an MPEG-4 decoder is required (no other software needs to be installed). The platform first checks the compatibility of the participant's device, as shown in Figure 11: the browser and OS used, the screen resolution (minimum required resolution of 1920×1080), and the page zoom level.

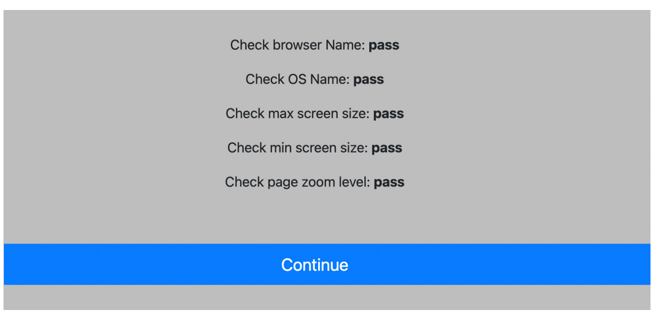


Fig. 11. Step 1: Verification of the compatibility of the participant's device.

Next, we ask for the participant's consent to collect and use their data (see Figure 12).

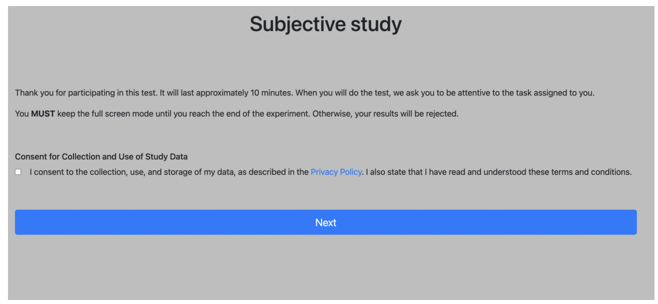


Fig. 12. Step 2: Participant's consent.

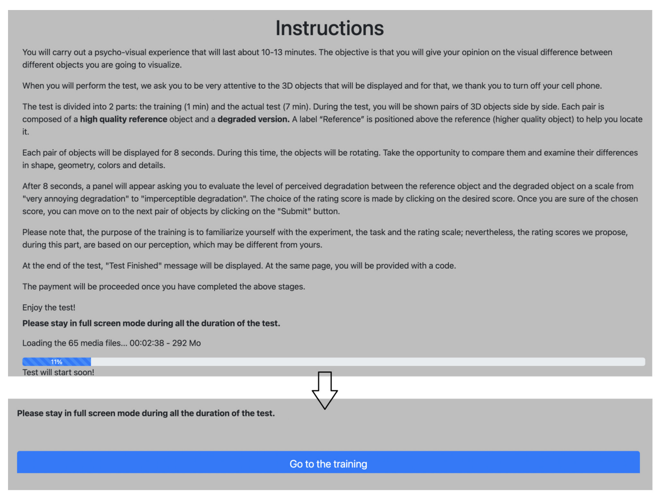


Fig. 13. Step 3: Experiment instructions.

The test instructions, shown in Figure 13, are then displayed to the participant with a progress bar, at the bottom of this page, showing the status of the loading process of all the video pairs that will be used in the test. This way, the videos of the source and distorted models are played simultaneously during the test, without any latency or unintended interruptions. When the loading is completed a start button appears leading to the training.

For the training, we selected 5 stimuli not included in the experiment and all referring to the same source model. Each stimulus represents one level of the five-level scale of the DSIS method. After displaying each pair of training videos for 8 sec, the rating interface is displayed for 5 sec and the proposed score assigned to this distortion is highlighted, as illustrated in Figure 14.

Once the training is completed the actual test began, see Figure 15. The pairs of videos (reference and distorted stimuli) are displayed side by side, in a random order to each participant. Participants cannot replay the videos or provide their score until the videos have been played completely. There is no time limit for voting and videos of the stimuli are not shown during that time.

[1] https://www.prolific.co/

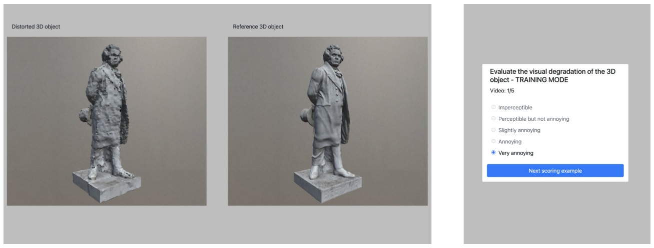


Fig. 14. Step 4: Training.

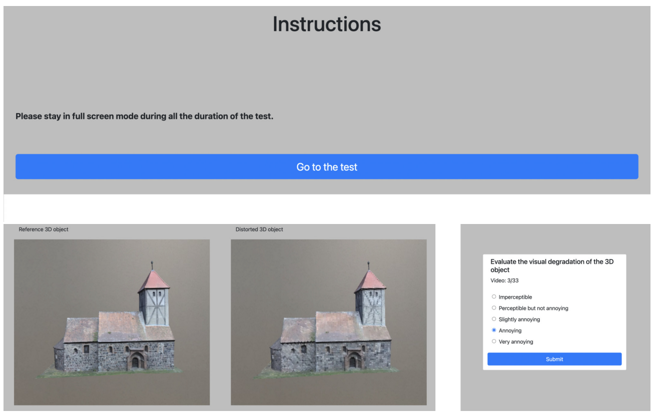


Fig. 15. Step 5: The experiment.

At the end of the experiment, participants will receive unique codes allowing them to get their remuneration, as shown in Figure 16.

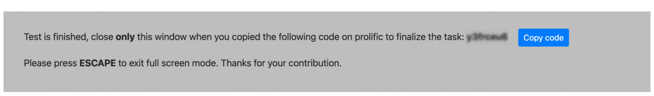


Fig. 16. Step 6: End of the experiment.

2.1 Pilot subjective experiment

Before conducting our large-scale subjective quality assessment experiment in crowdsourcing, we wanted to validate the experimental setup we implemented and study the number of participants needed in crowdsourcing to achieve the same accuracy (confidence intervals) as in a laboratory experiment. Thus, we conducted a pilot experiment with 30 stimuli selected from our dataset, using the rendering and experimental environment described in Section 4 of the paper. The stimuli were rated by 60 participants (i.e. 60 ratings collected per stimulus).

We computed the 95% Confidence Intervals (CIs) of the Mean Opinion Scores (MOSs) of the stimuli and assessed their evolution according to the number of ratings collected per stimulus (which is related to the number of participants involved in the test). Thus for each stimulus, we considered all possible combinations (without repetition) of N ratings and averaged the width of the CIs over all these ratings combinations. We compared the results to those obtained previously in a laboratory experiment conducted in Virtual Reality (VR) where 30 stimuli were evaluated by 30 participants. Results are shown in Figure 17.

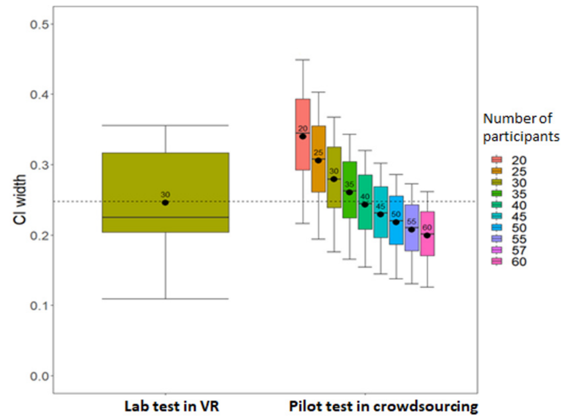
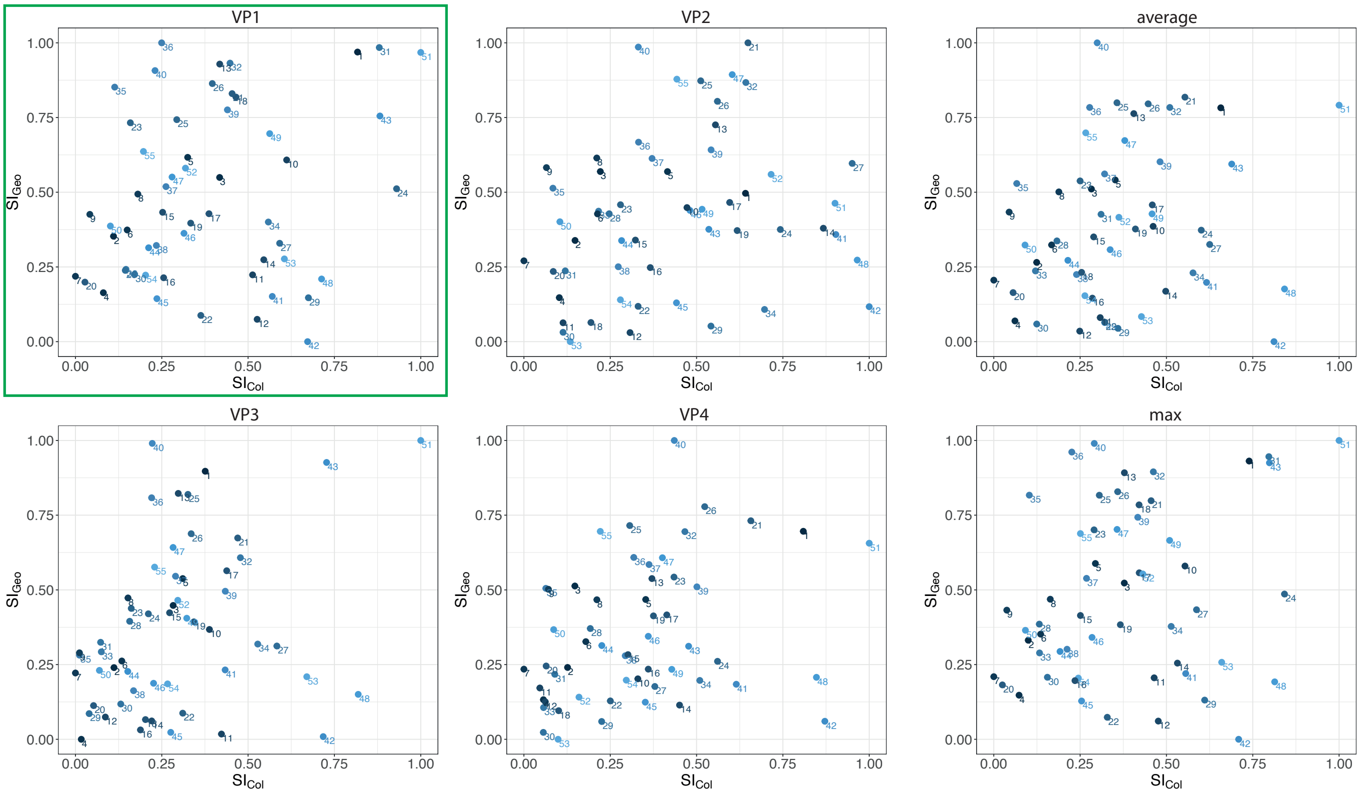


Fig. 17. Variation of Confidence Intervals (CIs) width according to the number of participants in the crowdsourcing and laboratory experiments.

The results show that almost 40 participants are required in the crowdourcing test to obtain the same accuracy (CIs) as the laboratory test. Keeping a margin for outliers, we considered having 45 scores per stimulus (i.e. each stimulus rated by at least 45 participants) in our large-scale crowdsourced experiment.

3 SETTINGS FOR IMAGE QUALITY METRICS

We compared our proposed metric Graphics-LPIPS to 3 state-of-the-art full-reference Image Quality Metrics (IQMs): *SSIM*, *HDR-VDP2*, *iCID*. For *SSIM*, we considered a local window of size 11×11 pixels. For the resolution used for *HDR-VDP2*, we considered 33.5 pixels per degree, which corresponds to the following experimental setting: stimuli presented on a calibrated 24" LCD display (resolution 1920×1200 pixel) at a constant viewing distance of 0.5m. The peak sensitivity parameter of *HDR-VDP2* was set to 2.4 and the selected output from this metric was the quality prediction Q. For the *iCID* metric, we considered the default parameters: equal weight of lightness, chroma, and hue. We computed the IQMs on 650 x 550 resolution snapshots taken from the main viewpoint of the stimuli.



a) each viewpoint

b) pooled over viewpoints

Fig. 18. a) We compute the geometric and semantic characterization on 4 different viewpoints regularly sampled around the bounding box of the object, the first viewpoint (circled in green) is the main viewpoint used in the paper. b) we pool the measures taken from the different views by using average (top) or max (bottom) pooling. The blue shade of the dot represents the id number of the object.

4 MESHES CHARACTERIZATION ON MULTIPLE VIEWS

We run our geometric and semantic characterization on 4 different viewpoints regularly sampled around the bounding box. The first viewpoint (VP1) corresponds to the main viewpoint of the model. The measures, normalized between 0 and 1, for each viewpoint are shown in Figure 18.a. In order to obtain a single score per mesh, we pooled the measures across the viewpoints by using either an average pooling or max-pooling (shown in Figure 18.b). Because the main viewpoint was chosen to be the most informative one, i.e. containing the maximum of information, using max-pooling on the 4 views leads to very similar results than using only this view. The proposed characterization strategy can thus be applied in both cases (automatic viewpoint sampling + max-pooling or manual viewpoint selection) with similar results.

5 ADDITIONAL EXPERIMENTS OF GRAPHICS-LPIPS

5.1 Evaluation on each individual fold

We evalute in Figure 19 the performance of *Graphics-LPIPS* and compares it to state-of-the-art Image Quality Metrics (IQMs), including the original LPIPS, on the test set of *each of our five folds* (each fold containing around 600 stimuli obtained from 11 source models). Similar to the aggregated results presented in the main paper, we show the performance of the metrics in terms of correlations and classification abilities.

We keep the first fold (#0) as our representative fold.

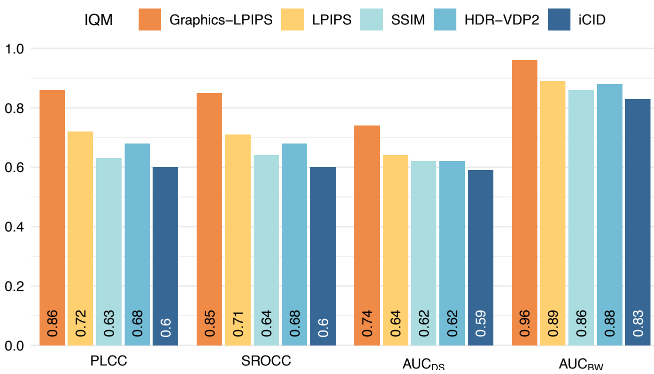
5.2 Patches pooling function

Our network first computes a similarity score for each patch. In order to produce a score for an entire image, we pool the scores for each patch of the image. We report in Table 1 the results using different pooling strategies: $L1$ (simple average), $L2$, $L3$ and max-pooling. The best results are obtained with the average pooling (L1), that we use in our final method.

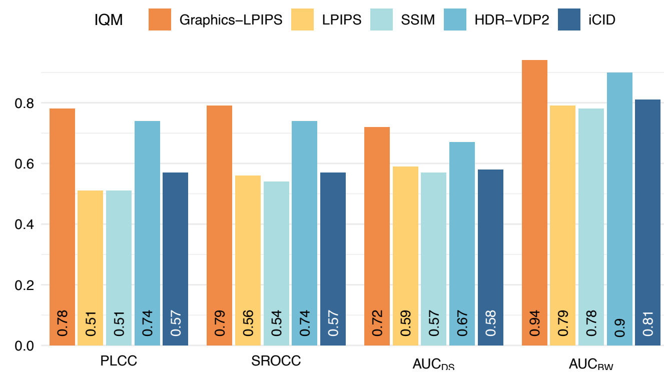
*(**Johanna:** give the formulas of Lp pooling?)*

Table 1. Performance comparison of different pooling strategies

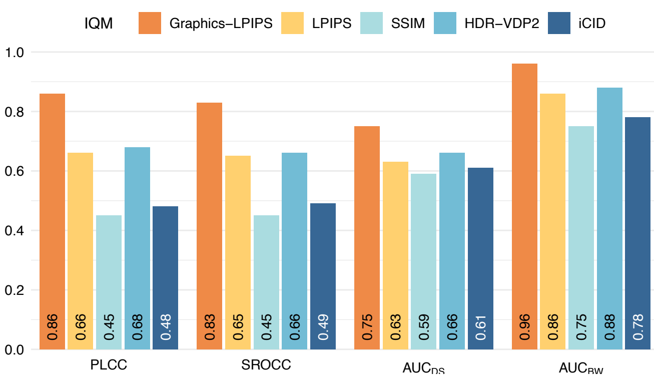
	L1 (average)	L2	L3	max
PLCC	**0.856**	0.838	0.812	0.819
SROCC	**0.845**	0.829	0.800	0.805



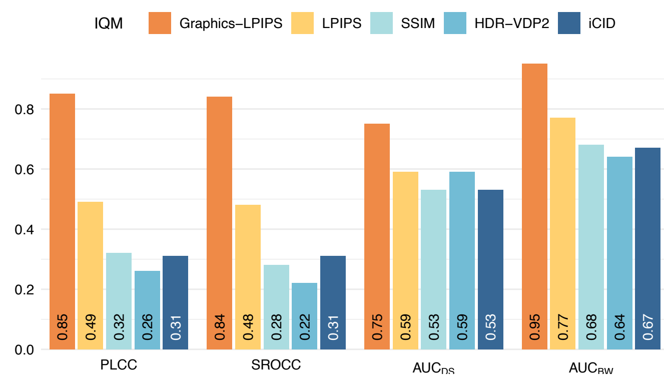
(a) Fold 0 (representative)



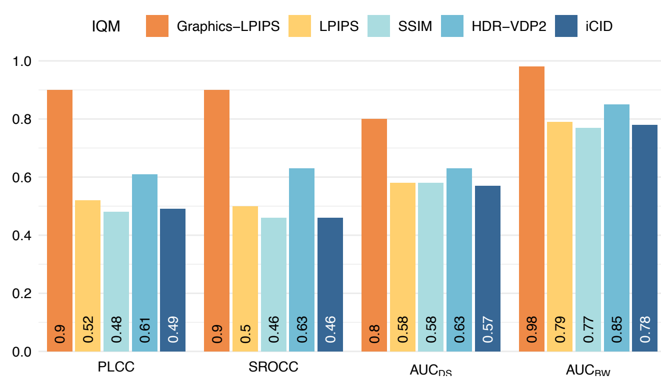
(b) Fold 1



(c) Fold 2



(d) Fold 3



(e) Fold 4

Fig. 19. Performances of our metric (Graphics-LPIPS) vs other Image Quality Metrics for each fold of our dataset.

6 APPLICATION

We used our metric Graphics-LPIPS to annotate the whole dataset of textured meshes and study the influence of several factors such as distortions and content characteristics on visual quality.

Indeed, we conducted a large-scale subjective experiment in crowdsourcing to evaluate the quality of a subset of 3000 stimuli carefully selected from over 343k. This subset of stimuli is associated with subjective scores and MOS values. To annotate the remaining stimuli of the dataset (over 340k), we applied Graphics-LPIPS to predict their MOSs. We referred to the predicted MOSs as pseudo-MOSs. Figure 20 illustrates the distribution of pseudo-MOSs for all stimuli in our dataset.

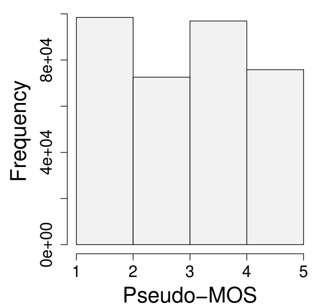


Fig. 20. Pseudo-MOSs distribution of all stimuli in the dataset.

6.1 ANOVA Table

The full ANOVA table about the influence of each distortion on perceived quality and their interactions (up to interactions between two factors) is reported in Table 2. All interactions are statistically significant.

6.2 Influence of distortion interactions on perceived quality

The impact of the combinations of the different distortions on the perceived quality differ from the cumulative impact of each distortion applied alone. The most visible and significant interactions are presented in Section 6.2 of the paper. In this section, we present other interesting distortion interactions impacting the perceived quality of textured meshes.

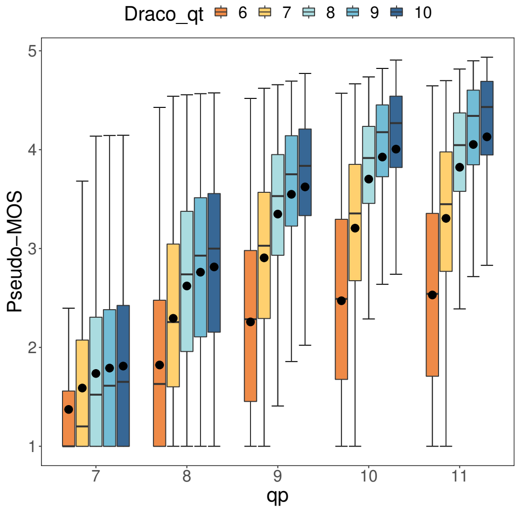
6.2.1 Interaction of geometry and texture coordinate quantization. It is interesting to observe that the perception of the distortion induced by the UV map quantization qt is affected by the quantization of the vertex positions qp. Figure 21a shows the interaction between these 2 factors. We can observe that for low qp values the improvement brought by increasing qt did not compensate the degradations generated by the strong geometric quantization and thus did not improve the MOSs much. Figure 21b shows 2 distorted versions of the bird (Model #33), both geometrically quantized with $qp = 6$. However, one stimulus has a higher qt ($qt = 10$) than the other ($qt = 6$). Both stimuli scored $MOS = 1$ (the lowest possible score); yet, the stimulus with less quantized texture coordinates ($qt = 10$) shows less degradation (see bird's eye and beak). This may be due to the five-level discrete categorical scale used in the DSIS method that does not allow for possible variations around best and worst qualities. We call this the "scale saturation effect".

Table 2. ANOVA table showing the influence of each distortion on perceived quality, and their interactions.

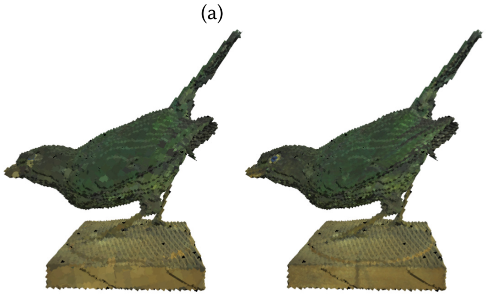
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
LoD_{simp}	9	1548	172.1	6577.178	< 2e-16 ***
qp	4	8414	2103.5	80411.662	< 2e-16 ***
qt	4	3242	810.5	30983.272	< 2e-16 ***
T_S	4	54	13.5	512.307	< 2e-16 ***
T_Q	4	200	50.1	1913.374	< 2e-16 ***
T_S:T_Q	16	26	1.6	62.824	< 2e-16 ***
T_S:qp	16	2	0.1	5.683	1.70e-12 ***
T_Q:qp	16	16	1.0	37.162	< 2e-16 ***
T_S:LoD_{simp}	36	4	0.1	4.225	3.25e-16 ***
T_Q:LoD_{simp}	36	4	0.1	4.493	< 2e-16 ***
qp:LoD_{simp}	36	1052	29.2	1117.201	< 2e-16 ***
T_S:qt	16	31	2.0	75.003	< 2e-16 ***
T_Q:qt	16	24	1.5	57.179	< 2e-16 ***
qp:qt	16	469	29.3	1120.438	< 2e-16 ***
LoD_{simp}:qt	36	182	5.0	192.871	< 2e-16 ***

Furthermore, looking at Figure 21a, it seems that the quantization of the model positions (qp) has more impact on the visual quality than the quantization of the UV map (qt): for low values of qp, we obtain a low MOS whatever the value of qt. Hence, we believe that for a given level of LoD_{simpL}, T_S and T_Q, the quality Q of a textured mesh can be represented by a multiplicative model as follows: $Q = Q_{qp}^{\alpha}.Q_{qt}^{\beta}$, where potentially $\alpha > \beta$.

6.2.2 Interaction of texture coordinate quantization and texture sub-sampling. The impact of the texture sub-sampling is strongly related to the mapping of the texture on the model surface. In fact, quantizing the texture coordinates with few bits ($qt \in \{6, 7, 8\}$) generates a "tiling effect", as illustrated in Figure 22. This effect is less visible on small textures. For instance, for $qt = 6$, stimuli with a texture size 512×512 scored better than those with a texture size 2048×2048. This is because sub-sampling the texture (reducing its size) reduces the high frequency information within the texture (which is like resampling using a low pass filter). Thus, the texture is smoothed, which decreases the tiling effect and therefore increases the MOS. qt and T_S are thus linked. These 2 parameters must be set with respect to each other: e.g., for low qt values (UV map strongly quantized), the texture size T_S must be decreased.



(a)

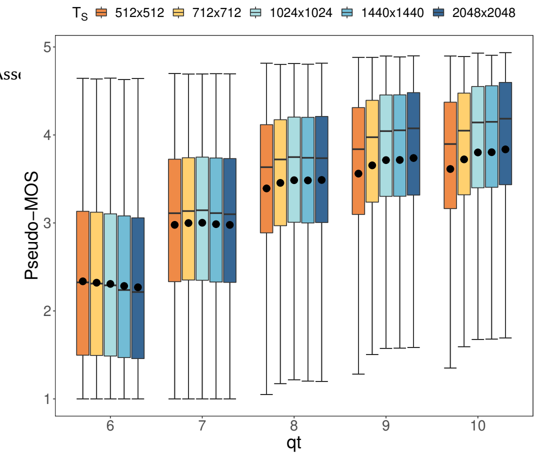


Model #33
$L1|\mathbf{7}|\mathbf{6}|2048 \times 2048|90$
Pseudo-MOS = 1

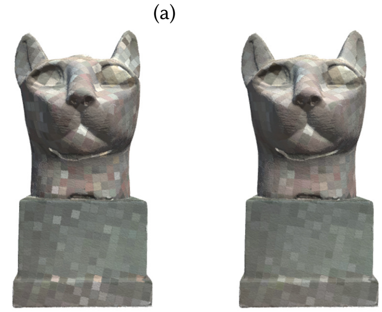
Model #33
$L1|\mathbf{7}|\mathbf{10}|2048 \times 2048|90$
Pseudo-MOS = 1

(b)

Fig. 21. (a) Boxplots of MOSs and (b) visual example illustrating the interaction between the geometry qp and texture coordinate qt quantization. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$. The impairments are less visible on the bird with the less quantified UV map (the one on the right $qt = 10$), yet both birds obtained the lowest possible score.



(a)



Model #45
$L1|11|\mathbf{6}|$
$2048 \times 2048|75$
Pseudo-MOS =
2.28

Model #45
$L1|11|\mathbf{6}|512 \times$
$512|75$
Pseudo-MOS =
2.76

(b)

Fig. 22. (a) Boxplots of MOSs and (b) visual example illustrating the interaction between the texture coordinate quantization qt and the texture sub-sampling T_S. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$. The UV map quantization artifacts ($qt = 6$) are less visible on the model with a small texture image (the one on the right) than on the one with a larger texture.

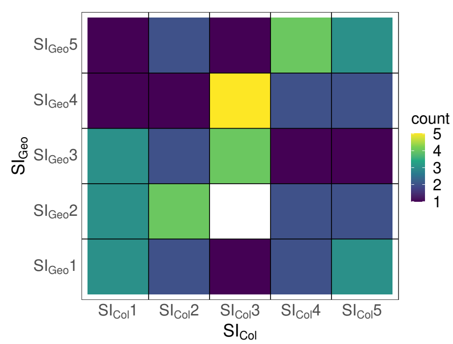


Fig. 23. Clusters of source models grouped by their geometric SI_{Geo} and color SI_{Col} characteristics.

6.3 Influence of content characteristics on perceived quality

We evaluated in our study the impact of model geometry and color complexity on the perception of distortions and thus on quality, using the content characterization measures (SI_{Geo} and SI_{Col}) described in Section 3.2 of the paper. The models were grouped into 5 clusters based on their geometric and color complexity: "$SI_{Geo}1$" contains the first 11 models with the least complex geometry, while "$SI_{Geo}5$" designates the 11 models with the most geometric details. Similarly, "$SI_{Col}1$" denotes the first 11 source models with the least color details, while "$SI_{Col}5$" refers to the models with the richest texture. Our clusters are well dispersed in the SI_{Geo}/SI_{Col} plane (cover a large range) as illustrated in Figure 23 which is an histogram representation of Figure 3.a. in the paper.

Figure 24 shows that for the same distortion parameters, the perceived quality is not the same: we obtained different ranges of MOS depending on the source models and their color and geometric characteristics.

(a) Model #48
$SI_{Geo}1$, $SI_{Col}5$
$L1|7|10|2048 \times 2048|90$
Pseudo-MOS = 2.76

(b) Model #22
$SI_{Geo}1$, $SI_{Col}3$
$L1|7|10|2048 \times 2048|90$
Pseudo-MOS = 1.68

(c) Model #4
$SI_{Geo}1$, $SI_{Col}1$
$L1|7|10|2048 \times 2048|90$
Pseudo-MOS = 1

(d) Model #23
$SI_{Geo}4$, $SI_{Col}1$
$L1|7|10|2048 \times 2048|90$
Pseudo-MOS = 3.9

Fig. 24. MOSs of different models with different geometric SI_{Geo} and color SI_{Col} characteristics and having undergone the same distortions ($LoD_{simpL}|qp|qt|T_S|T_Q$). For the same distortion parameters, the perceived quality was not the same: different ranges of MOS were obtained depending on the models' characteristics.

6.3.1 Influence of geometric complexity on the perception of texture coordinates quantization. To evaluate the influence of the color characteristics on the perception of degradations generated by the quantization of the texture coordinates (UV map) qt, we considered the subset of stimuli having a strongly quantized UV map ($qt \in \{6, 7, 8\}$) and the levels of all other distortions set at their best levels ($LoD_{simpL} \in \{L1, L2, L3\}$ & $qp \in \{10, 11\}$ & $T_Q \in \{75, 90\}$ & $T_S \in \{1440 \times 1440, 2048 \times 2048\}$)

Looking at Figure 25, we realize that the interaction between the geometry of the model and the quantization of the UV map is complex to evaluate, yet this interaction is significant (p-value $<< 0.0001$ according to ANOVA). Indeed, for low values of qt, the MOS decreases slightly from $SI_{Geo}1$ to $SI_{Geo}3$, then increases for $SI_{Geo}4$ and $SI_{Geo}5$. To better understand this behavior, we reported in Figure 26 visual examples of models $\in \{SI_{Geo}4, SI_{Geo}5\}$. We noticed that the MOS values are not systematically high for all these models. It depends on the models, specifically the texture seams and the quality of the surface parameterization: i.e., the fragmentation of the texture atlas and the quality of the atlas packing. Quantization artifacts are clearly more visible on models whose texture atlas is highly fragmented (high number of texture seams) and/or not efficiently packed (see Model #1 in Figure 26). In contrast, UV quantization artifacts are less visible for models having homogeneous/uniform texture colors and/or less fragmented textures (low number of texture seams), as can be seen for Model #31 in Figure 26.

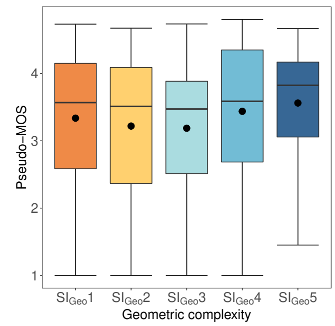


Fig. 25. Boxplots of the MOSs illustrating the influence of the geometric complexity SI_{Geo} of the models on the perceived degradation of texture coordinates quantization qt.

Model #1
$\mathbf{SI_{Geo}}5$, $SI_{Col}5$
$L1|11|\mathbf{6}|2048 \times 2048|90$
Pseudo-MOS = 1.6

Model #31
$\mathbf{SI_{Geo}}5$, $SI_{Col}5$
$L1|11|\mathbf{6}|2048 \times 2048|90$
Pseudo-MOS = 3.97

Fig. 26. Visual examples illustrating the impact of texture coordinates quantization on the perceived quality of textured meshes. Models are presented with their texture seams highlighted and their texture map. Acronyms refer to the following combination of distortion parameters: $LoD_{simpL}|qp|qt|T_S|T_Q$. The UV map quantization artifacts ($qt = 6$) are more visible on Model #1 which has a larger number of texture seams than Model #31.