EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 43 (2024), Number 2

SUPPLEMENTAL MATERIAL - Predicting Perceived Gloss: Do Weak Labels Suffice?

Julia Guerrero-Viu[1*§], J. Daniel Subias[1*], Ana Serrano[1], Katherine R. Storrs[2], Roland W. Fleming[3,4],
Belen Masia[1] & Diego Gutierrez[1]

[1]Universidad de Zaragoza, I3A, Spain [2]University of Auckland, New Zealand [3]Justus Liebig University Giessen, Germany
[4]Center for Mind, Brain and Behavior, Universities of Marburg and Giessen, Germany
[*]Joint first authors §Corresponding author: juliagviu@unizar.es

The supplemental material of this paper includes:

- Our training and test datasets, gloss prediction models and code, available in `https://graphics.unizar.es/projects/perceived_gloss_2024/`.
- This pdf document, offering additional information and details on the following topics:
  - (S1) Datasets: Additional Details
  - (S2) Additional Results

S1. Datasets: Additional Details

S1.1. Training Dataset

Our analytical dataset comprises single-object scenes with fifteen different geometries. The dataset consists of eight geometries from the *Serrano* dataset and seven newly introduced geometries [Thi] (see Figure 1, left). Additionally, we have incorporated seventeen real-world illuminations, including nine illuminations from the *Serrano* dataset and eight new illuminations [Pol] (see Figure 1, right).

For each combination of illumination and geometry, we vary the roughness r and specular s parameters of Disney's Principled BSDF [BS12, Bur15]. The roughness values vary from 0.1 to 5.0 ($\{0.1, 0.15, 0.24, 0.36, 0.57, 0.87, 1.36, 2.1, 3.24, 5.0\}$) and the specular values vary from 0.0 to 0.5 ($\{0.0, 0.125, 0.25, 0.375, 0.5\}$).

Additionally, we generate three random colors for each combination of roughness and specular values. The colors are sampled from the perceptually uniform HSLuv color space. Therefore, our analytical dataset consists of a total of 38,250 images (15 geometries × 17 illuminations × 10 roughness values × 5 specular values × 3 random colors). To further sample the objects' surface, we randomly rotate the geometry on each axis between -15º and 15º for each image, as shown in Figure 2.

S1.2. Test Dataset

Our test dataset includes twenty baseline single-object scenes. These baseline images include two illuminations, two geometries, and five measured materials from the MERL dataset [MPBM03] with distinguishable gloss levels, as explained in the main document.

We expand our test dataset by variations in different confounding factors: three additional illuminations, four additional rotations, four additional levels of geometry complexity, and five different levels of the specular parameter for the analytical fittings of the materials. Therefore we have 290 additional scenes: 30 from variations in illumination (3 extra illuminations × 2 geometries × 5 measured materials), 80 from variations in rotation (2 baseline illuminations × 2 geometries × 5 measured materials × 4 extra rotations for the object), 80 from variations in the level of geometry complexity (2 baseline illuminations × 2 geometries × 5 measured materials × 4 extra levels of geometry complexity) and 100 from variations on the level of the specular parameter for the analytical fittings of the materials (2 baseline illuminations × 2 geometries × 5 analytical fittings of the materials × 5 levels of the specular parameter). In total, our test dataset comprises 310 images with a balanced distribution of human ratings in comparison with the test set A from Serrano et al. [SCW*21] (with only one geometry and one illumination), as shown in Figure 3.

Annotation Process

During the annotation process of our test dataset, the five subjects were shown four example images with different gloss levels as a training phase to calibrate what is the expected range of glossiness in our test dataset. Then, each annotator rated all the images in our test dataset in the 7-point Likert scale, and in a random order, without repeating the stimuli. At all times during the annotation process, our interface also included a small "info" button with two example images of the extreme gloss-level anchors [Cun21]. Figure 4 shows a screenshot of the perceptual study, as seen by the participants. The stimuli is shown on the left part of the screen while the list of ratings (from 1 to 7) is shown on the right. All annotators were exposed once to each image in our test dataset at 1024x1024 resolution, in an SDR display and fixed lighting.

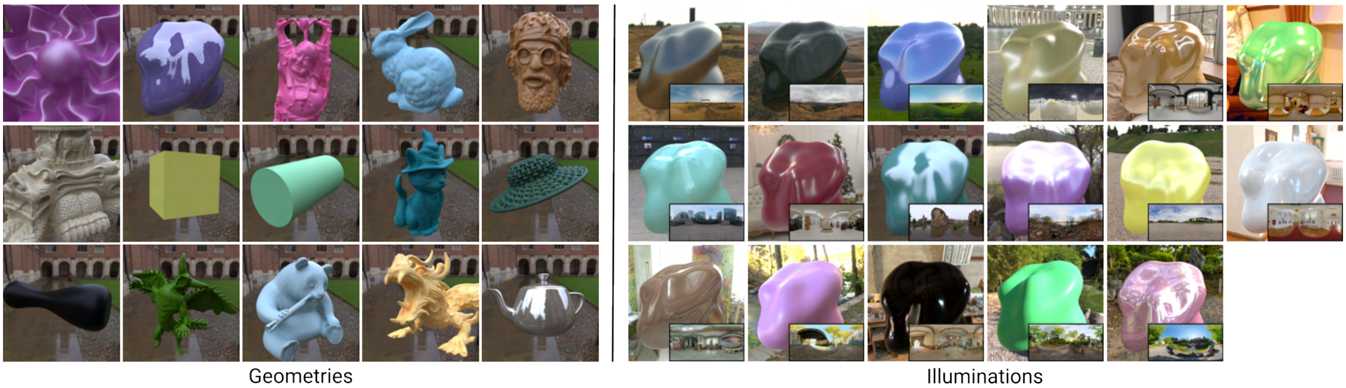


Figure 1: *Left: The 15 geometries included in our training dataset under the "cambridge" illumination and with randomly-colored analytical materials. Right: The 17 illuminations present in our training dataset and corresponding rendered blobs with randomly-colored analytical materials.*

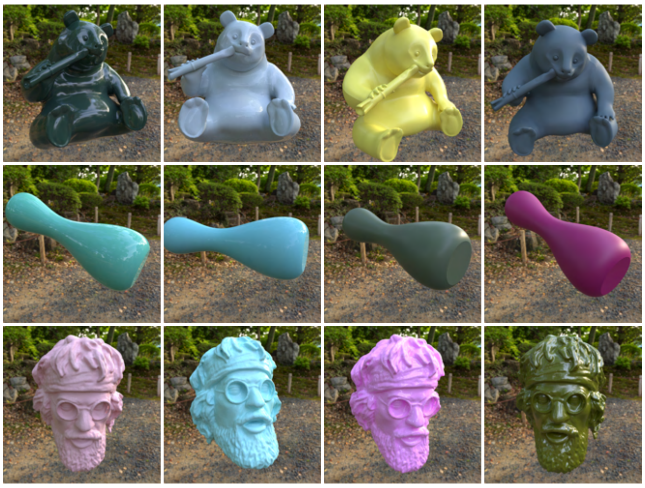


Figure 2: *Examples of 4 random rotations on each axis between -15º and 15º for the "panda", "handler" and "chong_head" geometries under the same illumination with randomly-colored analytical materials.*

S2. Additional Results

S2.1. Weakly Supervised Learning

In Figure 5, we show the distribution and Pearson correlations of the ground-truth data (human annotations) on our test dataset with respect to the predictions of our weakly supervised gloss predictors trained on the following data: using only strong human labels (with the 100% of the Serrano dataset or with only 20% of it), combining these strong labels with our weak labels based on either BSDF model, image statistics or industry metrics, and using only our weak labels.

Additionally, we note that combining human annotations with our weak labels is not equivalent to combining them with simply noisy annotations. To show it, we train a model with the 20% of the Serrano dataset jointly with 80% of the Serrano dataset with jit-

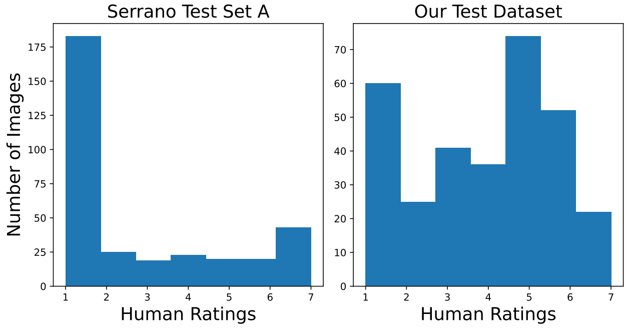
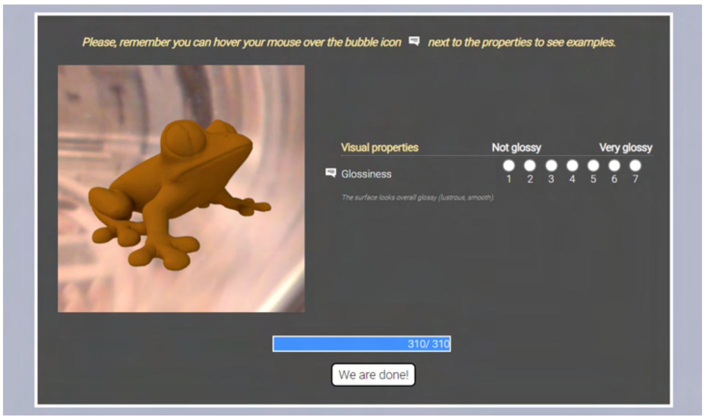


Figure 3: *Histogram of the human gloss ratings (range $[1,7]$) in test set A from Serrano et al. [SCW*21] (left) and our test dataset (right). Our dataset is more evenly distributed, which enables a more complete evaluation of gloss predictions, specially for more glossy appearances.*



Please, remember you can hover your mouse over the bubble icon next to the properties to see examples.

The surface looks overall glossy (lustrous, smooth)

Figure 4: *Screenshot of the user study as seen by the annotators. Stimuli is shown on the left, the annotators have to select a rating for the gloss level on the right. On the top there is a "info" button to show two example images of the extreme gloss-level anchors.*

ter (additive Gaussian noise with mean=0.0 and std=1.0), which obtains MAE=0.18 on our test data set, less accurate than our weakly supervised models trained with S.20%+weak labels (e.g., MAE=0.15 for S.20%+BSDF, see Table 1 in the main paper).

We provide additional qualitative results of our weakly supervised gloss predictors trained with strong labels of the *Serrano* dataset jointly with our three different weak labels: based on the BSDF model (Figure 6), based on image statistics (Figure 7) and based on industry metrics (Figure 8).

S2.2. Consistency Evaluation

In Figure 9 we show qualitative results of our weakly supervised gloss predictors trained with our BSDF weak labels for variations in the level of geometry complexity (i.e. bumpiness of the surface) and variations in illumination, from our test dataset. The first and second rows show the results when increasing the level of geometry complexity: The human perception of gloss (GT) remains constant when the material is diffuse; when the material is highly specular, human perception of gloss tends to decrease if the geometry is very complex. Our predictor successfully follows this behavior in its predictions. The third and fourth rows show the results across all illuminations in our test dataset. As expected, *glacier* illumination decreases the human perception and predictions of gloss as it causes the materials to appear highly diffuse. In Figure 10, we also show additional results for variations in rotations and the specular parameter.

References

[BS12] BURLEY B., STUDIOS W. D. A.: Physically-based shading at Disney. In *ACM SIGGRAPH* (2012), vol. 2012, pp. 1–7.

[Bur15] BURLEY B.: Extending the disney brdf to a bsdf with integrated subsurface scattering. *ACM SIGGRAPH Course: Physically Based Shading in Theory and Practice 19* (2015).

[Cun21] CUNNINGHAM A.: The united states and its obligations under the optional protocol to the convention on the rights of the child on the sale of children, child prostitution and child pornography to combat child exploitation in the digital world. *Ga. J. Int'l & Comp. L. 50* (2021), 670.

[MPBM03] MATUSIK W., PFISTER H., BRAND M., MCMILLAN L.: A data-driven reflectance model. *ACM Transactions on Graphics 22*, 3 (2003), 759–769.

[Pol] Poly Haven: HDRIs. https://polyhaven.com/hdris. [Last accessed 28.09.2023].

[SCW*21] SERRANO A., CHEN B., WANG C., PIOVARČI M., SEIDEL H.-P., DIDYK P., MYSZKOWSKI K.: The effect of shape and illumination on material perception: model and applications. *ACM Transactions on Graphics 40*, 4 (2021), 1–16.

[Thi] Thingiverse. https://www.thingiverse.com/. [Last accessed 28.09.2023].

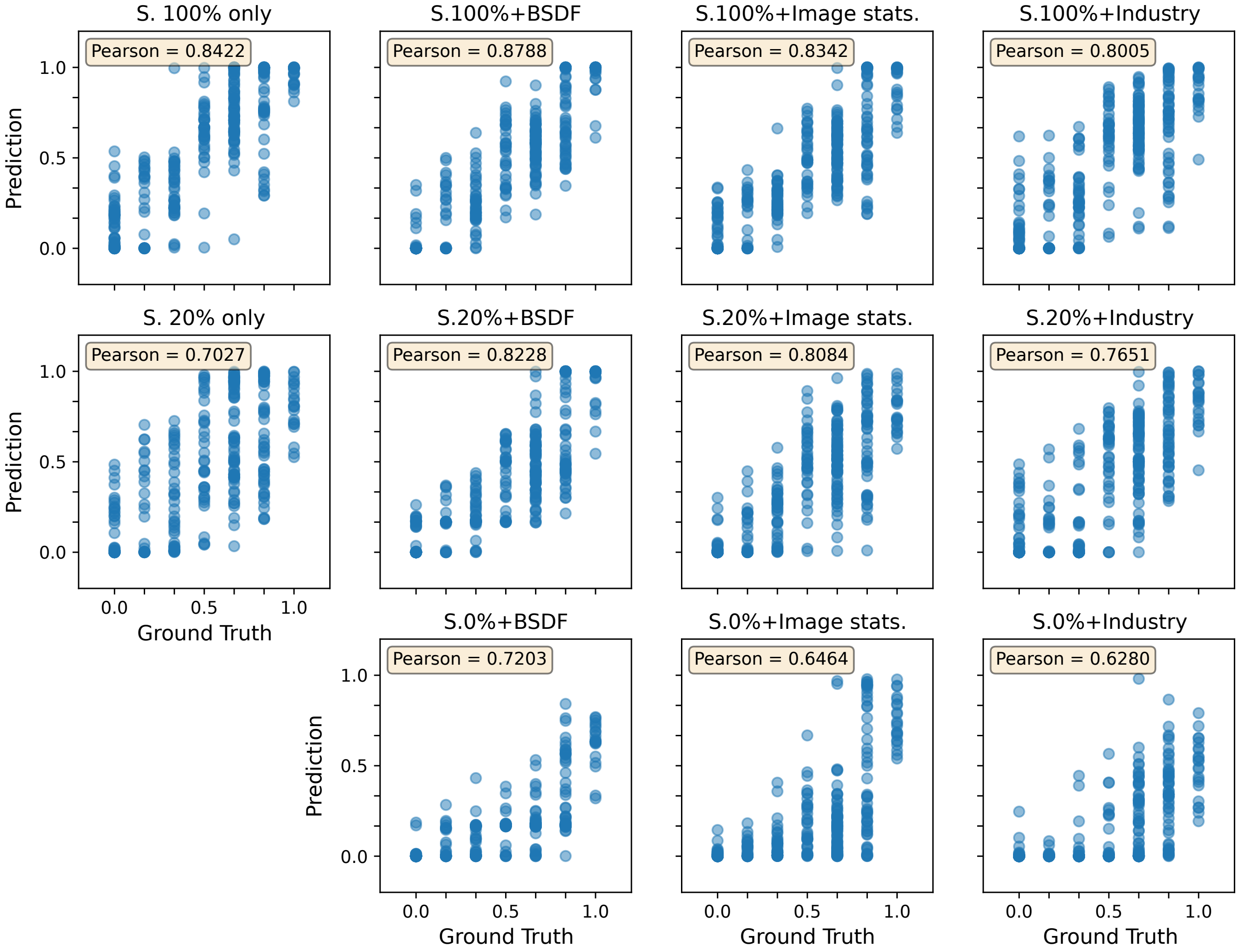


Figure 5: *Pearson correlations and visual distribution of the ground-truth data (x-axis) vs. the predictions of our gloss predictor (y-axis) trained on the following data: using the 100% of the Serrano dataset jointly with our weak labels (first row), using the 20% of the Serrano dataset jointly with our weak labels (second row) and using only our weak labels (third row). All ratings are in range* $[0,1]$. *Notice that the ground truth is discrete, because it was obtained from normalizing the Likert-scale ratings in range* $[1,7]$ *to range* $[0,1]$.

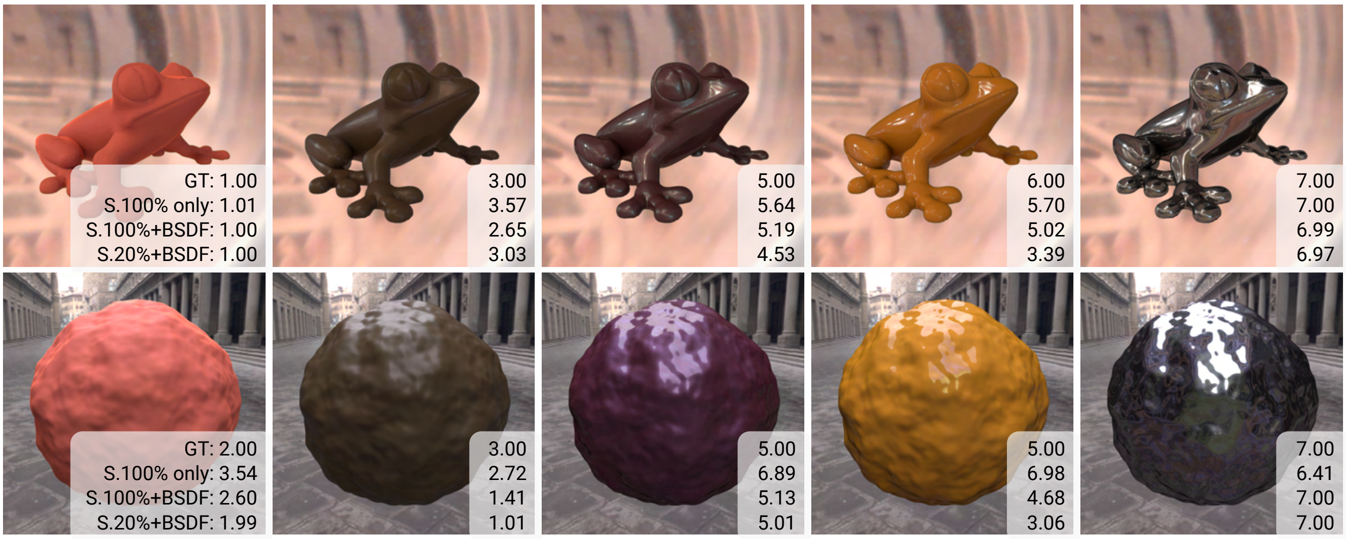


Figure 6: *Qualitative results of our weakly supervised gloss predictor trained with our BSDF weak labels jointly with the 100% of the Serrano dataset (S.100%+BSDF) and our weakly supervised gloss predictor trained with our BSDF weak labels jointly with the 20% of the Serrano dataset (S.20%+BSDF). The numbers in the insets display also the ground-truth judgements (GT) and predictions from our supervised model trained on the Serrano dataset only (S.100% only). All gloss ratings are in range* $[1,7]$.

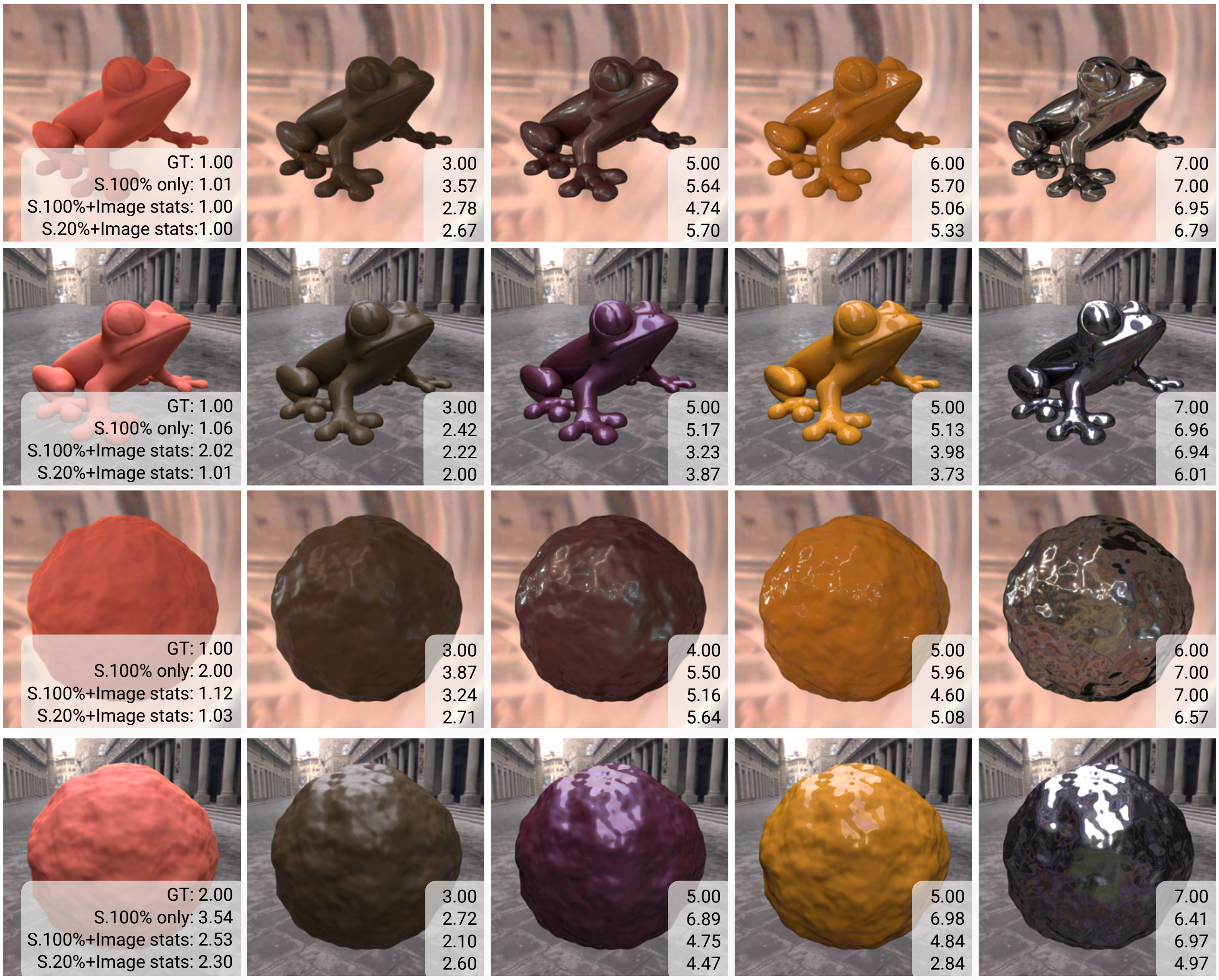


Figure 7: *Qualitative results of our weakly supervised gloss predictor trained with our weak labels based on image statistics jointly with the 100% of the Serrano dataset (S.100%+Image stats.) and our weakly supervised gloss predictor trained with our weak labels based on image statistics jointly with the 20% of the Serrano dataset (S.20%+Image stats.). The numbers in the insets display also the ground-truth judgements (GT) and predictions from our supervised model trained on the Serrano dataset only (S.100% only). All gloss ratings are in range* $[1,7]$.

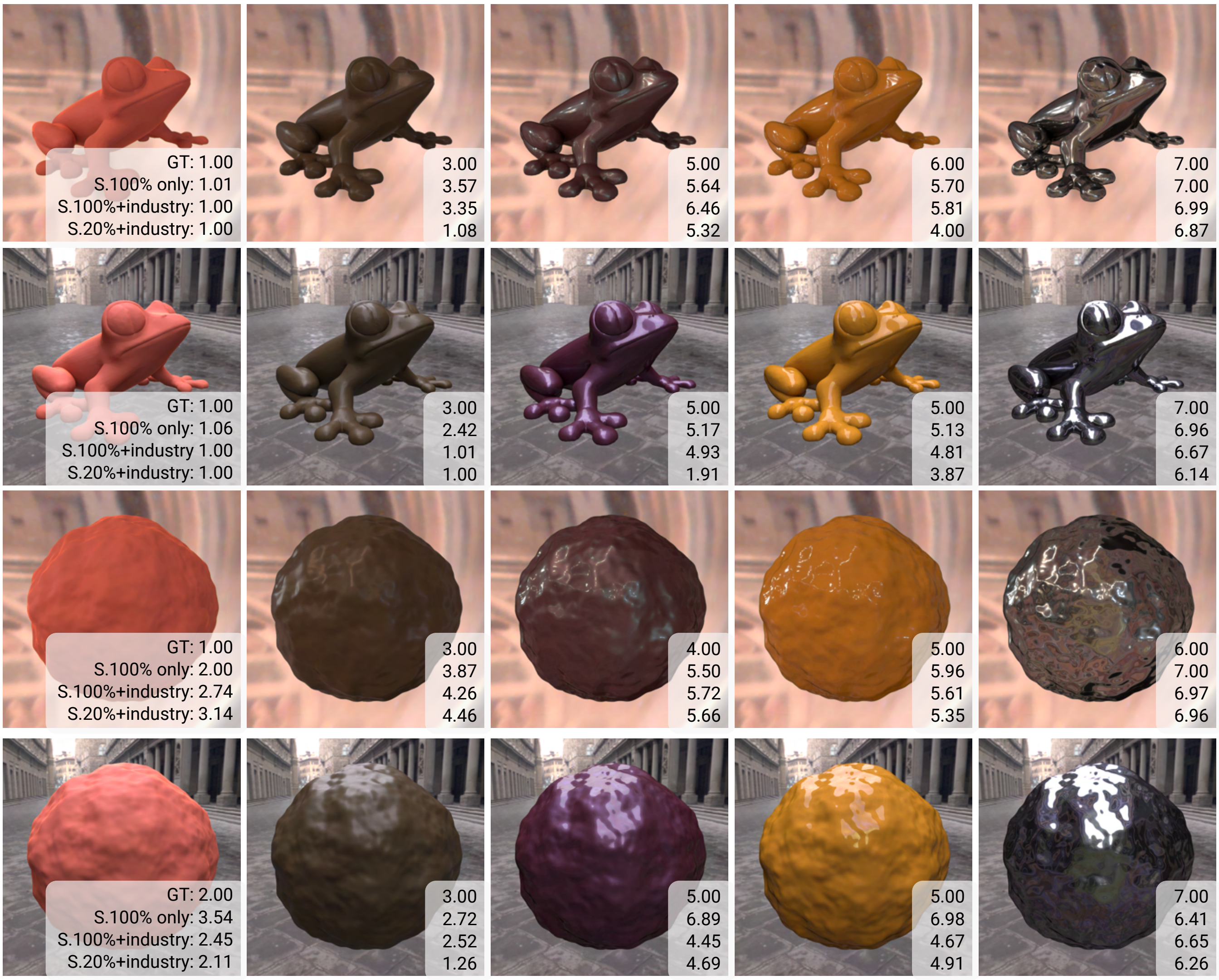


Figure 8: *Qualitative results of our weakly supervised gloss predictor trained with our weak labels based on industry metrics jointly with the 100% of the Serrano dataset (S.100%+Industry) and our weakly supervised gloss predictor trained with our weak labels based on industry metrics jointly with the 20% of the Serrano dataset (S.20%+Industry). The numbers in the insets display also the ground-truth judgements (GT) and predictions from our supervised model trained on the Serrano dataset only (S.100% only). All gloss ratings are in range* $[1,7]$.

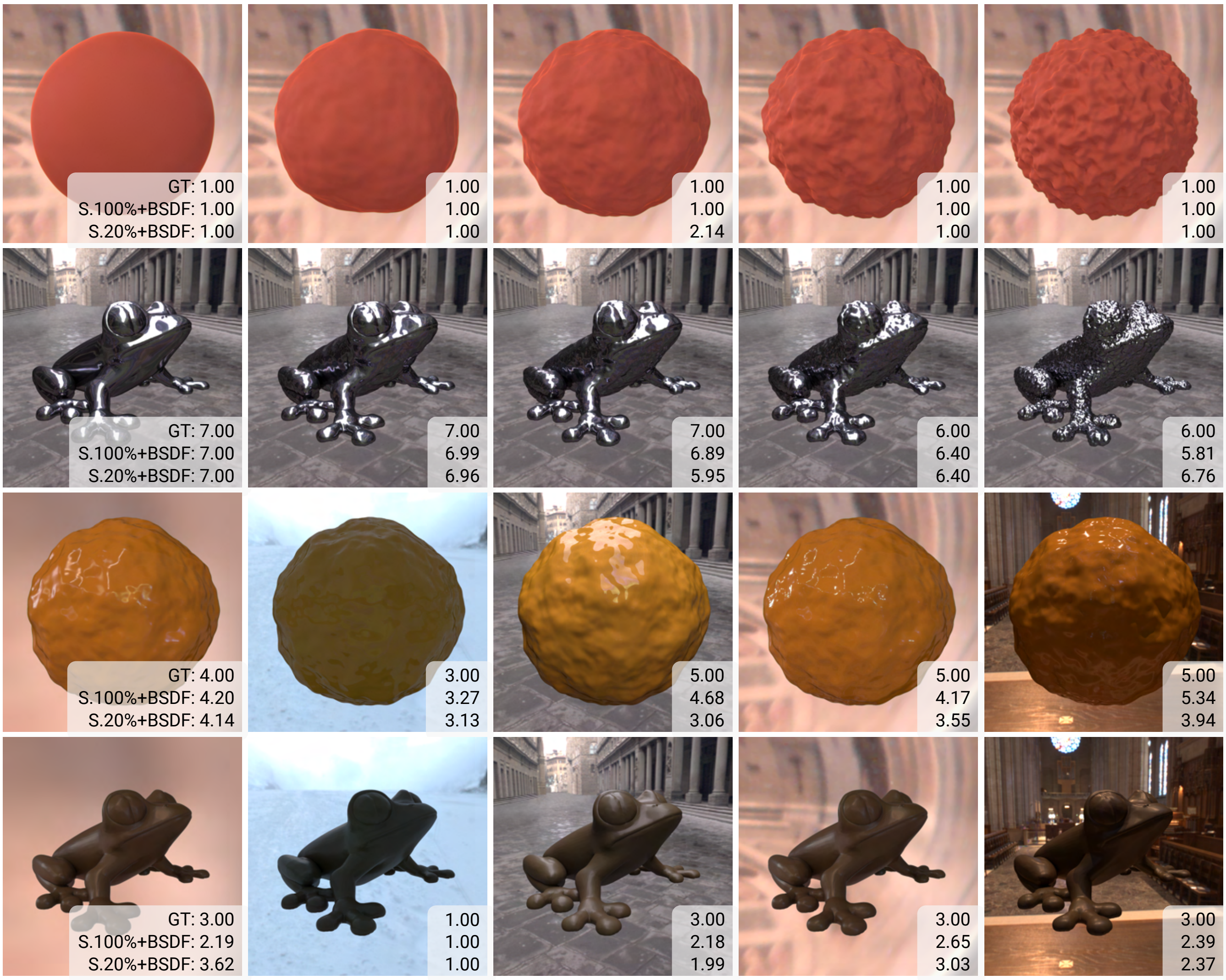


Figure 9: *Qualitative results of our weakly supervised gloss predictor trained on the 100% of the Serrano dataset jointly with our BSDF weak labels (S.100%+BSDF) and our weakly supervised gloss predictor trained on the 20% of the Serrano dataset jointly with our BSDF weak labels (S.20%+BSDF), when varying one confounding factor on our test dataset. The numbers in the insets display also the ground-truth judgements (GT). We show: variation across different levels of geometry complexity for the "bumpy_sphere" geometry with "st_peters" illumination and "pink_plastic" material, variation across different levels of geometry complexity for the "frog" geometry with "uffizi" illumination and "aluminium" material, variation across the five illuminations in our test dataset for baseline "bumpy_sphere" geometry with the "specular_yellow_phenolic" material and variation across the five illuminations in our test dataset for baseline "frog" geometry with the "fruitwood" material. All gloss ratings are in range* $[1,7]$.

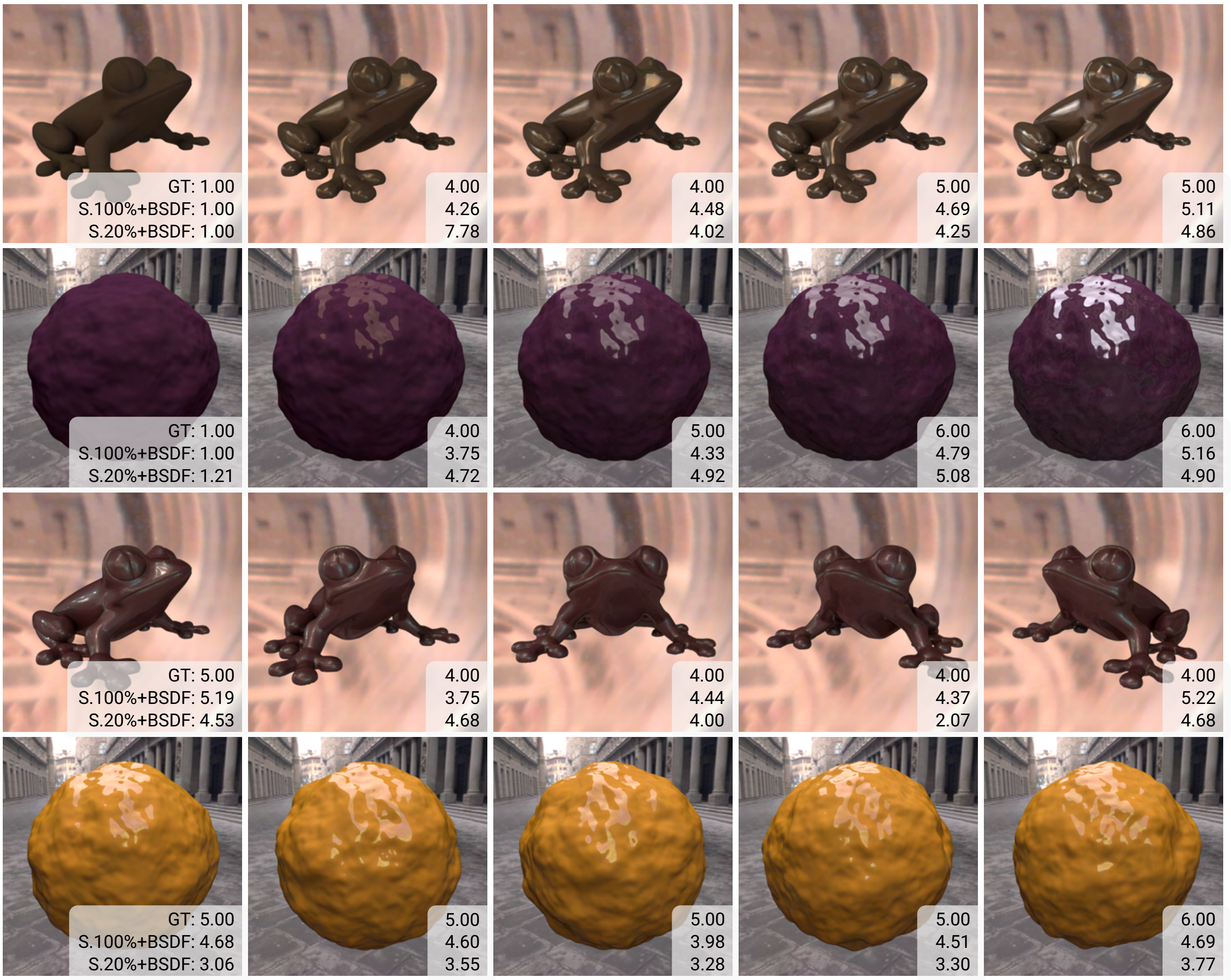


Figure 10: *Qualitative results of our weakly supervised gloss predictor trained on the 100% of the Serrano dataset jointly with our BSDF weak labels (S.100%+BSDF) and our weakly supervised gloss predictor trained on the 20% of the Serrano dataset jointly with our BSDF weak labels (S.20%+BSDF), when varying one confounding factor on our test dataset. The numbers in the insets display also the ground-truth judgements (GT). We show: variation across increasing specularity for the "frog" geometry with "st_peters" illumination and the Ward-Duer BRDF fitting of "fruitwood" material, variation across increasing specularity for the "bumpy_sphere" geometry with "uffizi" illumination and the Ward-Duer BRDF fitting of "violet_acrylic" material, variation across different rotations for the "frog" geometry with "st_peters" illumination and "violet_acrylic" material and variation across different rotations for the "bumpy_sphere" geometry with "uffizi" illumination and "specular_yellow_phenolic" material. All gloss ratings are in range* $[1,7]$.