DOI: 10.1111/cgf.15037
EUROGRAPHICS 2024 / A. Bermano and E. Kalogerakis
(Guest Editors)

COMPUTER GRAPHICS *forum*

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

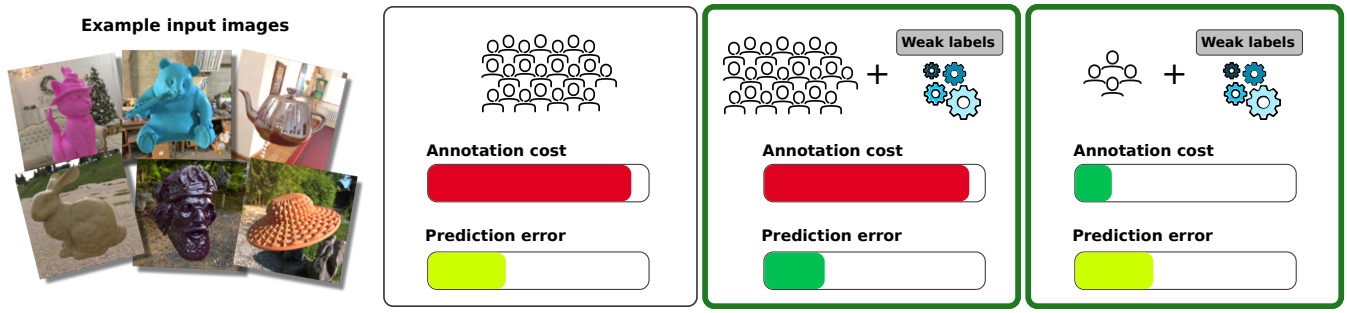

Figure 1: *We explore the benefits of weakly supervised learning to predict perceived gloss from images. Commonly used supervised models require costly manual annotations of large datasets (left). In contrast, we propose to leverage simple weak labels, which can be automatically computed, achieving a twofold contribution: (1) our weak labels can be effectively combined with such costly annotations to reduce gloss prediction error, if accuracy is the priority (center); and (2) they enable a significant reduction in manual annotations of up to 80% without sacrificing accuracy, if efficiency is the priority (right).*

Abstract

Estimating perceptual attributes of materials directly from images is a challenging task due to their complex, not fully-understood interactions with external factors, such as geometry and lighting. Supervised deep learning models have recently been shown to outperform traditional approaches, but rely on large datasets of human-annotated images for accurate perception predictions. Obtaining reliable annotations is a costly endeavor, aggravated by the limited ability of these models to generalise to different aspects of appearance. In this work, we show how a much smaller set of human annotations ("strong labels") can be effectively augmented with automatically derived "weak labels" in the context of learning a low-dimensional image-computable gloss metric. We evaluate three alternative weak labels for predicting human gloss perception from limited annotated data. Incorporating weak labels enhances our gloss prediction beyond the current state of the art. Moreover, it enables a substantial reduction in human annotation costs without sacrificing accuracy, whether working with rendered images or real photographs.

CCS Concepts

• ***Computing methodologies*** → ***Perception; Dimensionality reduction and manifold learning; Supervised learning;***

1. Introduction

The advent of powerful generative deep learning models for image synthesis and editing creates new opportunities, as well as new challenges, for computer graphics. In this context, techniques to navigate the massively high-dimensional latent spaces of such models in ways that are perceptually meaningful to humans become increasingly vital. Within the field of computer graphics, understanding and modeling human perception of material appearance has long been a fundamental challenge. This understanding is essential for accurately simulating the visual aspects of the physical world as perceived by humans and it has potential applications ranging from design and virtual prototyping to image editing and fabrication processes [CPW*22, SL23]. In particular, gloss is one of the most perceptually salient appearance at-

tributes [CK15, Fle17]. Unfortunately, how glossy a surface appears to a human viewer is a product of complex and only partially understood interactions between surface reflectance, illumination, and geometry [Fle14, LSGM21]. Perceived gloss, despite being grounded in photogeometric features such as the properties of reflected highlights [MKA12], cannot be well captured by either linguistic descriptors or objective measures, such as surface reflectance. This makes gloss an excellent candidate for a learned perceptual metric.

Recently, it has been shown that complex representations of visual inputs are likely needed to capture perception of gloss in images [DLG*20, FS19, LSGM21], drawing attention to deep learning approaches. Gloss perception is a high-dimensional problem that cannot be easily modeled by combining only physical material properties like roughness or specularity [Fle14, CK15]. Such a model could not, for instance, account for the effects of shape or illumination on perceived gloss [MKA12, SAF21, SCW*21]. Simple image statistics also fall short, due to the importance of spatial information (e.g., congruence of highlights, lowlights, and shading). Consequently, deep learning offers a framework for computing suitably complex visual features from which perceptual gloss metrics may be directly learning from detailed human annotations [LMS*19, SCW*21]. State-of-the-art gloss predictions have been achieved using supervised methods, as demonstrated by Serrano et al. [SCW*21]. However, these approaches heavily rely on a substantial amount of human-annotated images, which can be prohibitively expensive to obtain. For instance, Serrano et al. [SCW*21] collected over 200,000 gloss ratings from more than 3,000 human participants for their dataset. Alternatively, methods employing fully unsupervised models, such as the one proposed by Storrs et al. [SAF21], face challenges in capturing material properties in complex stimuli and realistic images. Object appearance in realistic images is influenced by multiple factors, including geometry, lighting, and viewpoint, creating a prohibitively high-dimensional space to learn in a fully unsupervised fashion from limited training data. Moreover, even after training, unsupervised models require further adjustments to align their latent dimensions with human perception.

In this work, we show that costly human annotations (*strong* labels) can be effectively combined with automatically derived *weak* labels in the context of learning a low-dimensional image-computable gloss metric. First, we demonstrate that our weakly supervised approach achieves higher gloss prediction accuracy than previous supervised methods (Figure 1, center). It is not trivial that such an approach should work – noisy weak labels can in some domains "dilute" high-quality strong labels, reducing a model's performance even as more training data is provided [LGZ19]. Then, we leverage our weak labels to *reduce* the annotation cost, allowing for an 80% reduction in the amount of human-labeled data needed, without losing accuracy (Figure 1, right). In addition, we identify the need for a *controlled* dataset to systematically evaluate gloss prediction models, as opposed to crowd-sourced existing datasets. For this purpose, we create a new test dataset that includes systematic variations in rotation, geometry complexity, illumination and specularity, and covers a reasonable and balanced range of glossiness levels. This new test dataset is annotated under constant, controlled viewing conditions that enable a *reliable* evaluation of gloss prediction models. Finally, we show how our weakly supervised gloss predictors are consistent with human perception for such systematic variations and are able to generalize reasonably to "in-the-wild" real images.

In summary, we present the following contributions:

- A study of three possible automatic weak labels (simple and cheap to obtain) that can effectively be leveraged to predict human perception of gloss without the need for a large amount of costly annotated data.
- A test dataset with controlled variations and reliable annotations to evaluate perceptual gloss prediction models.
- An accurate gloss predictor that outperforms previous methods, is consistent across changes in object view or illumination, and generalizes to real-world photographs.

Our datasets and gloss prediction models, as well as the training and evaluation code, are available in `https://graphics.unizar.es/projects/perceived_gloss_2024/`.

2. Related Work

Weakly supervised learning [Zho18] has been widely applied in computer vision and graphics in several domains such as image classification [XXY*15, CZZ*20], object detection [RYY*20, CFJ*20], semantic segmentation [BRFFF16, LY20, KJHH23], or image retargeting [CPO*17]. However, in the fields of material appearance and perception science, contemporary studies still rely on tedious and expensive human annotations [LMS*19, DLC*22], which are the standard in psychophysical experiments. In this paper, we perform the first study of weakly supervised learning focusing on the field of material appearance, particularly on perceptual gloss prediction.

2.1. Gloss Metrics

Finding an objective measure of gloss is a long-standing problem in computer graphics, industry, perception research and related fields. Since the work from Hunter et al. [H*37], it has been recognized that a single physical measure is not sufficient to quantify physical gloss. An active field of research on the perceptual aspects of gloss has demonstrated that perceived gloss depends not only on surface reflectance properties but also on geometry, illumination, and motion [FDA01, FDA03, PFG00, WAKB09]. Chadwick et al. [CK15] provides a review of the study of gloss perception and measurement.

One intuitive approach to quantifying gloss would be to consider a physical parameter (e.g., roughness) from an analytical BSDF model, as an objective measurement. However, such parameters of analytical models do not correlate well with perceived gloss [NDM06, WAKB09]. Moreover, it is difficult to adjust BSDF parameters to approximate real-world measured materials while maintaining their perceptual quality [LBFS21]. Alternatively, there are works showing that certain image statistics [MNSA07], or visual cues [MKA12] correlate with perceptual judgments of gloss, at least for simple stimuli. Last, the work of Westlund and Meyer [WM01] explores the objective measurement

of physical gloss on analytical BSDF models following industry standards [HJ39].

Following each of these proposals, we explore three different objective metrics to label gloss perception of rendered images, based on BSDF parameters, image statistics, and industry standards, and analyze their usefulness as weak labels to guide learning-based algorithms for gloss prediction.

2.2. Dimensionality Reduction for Material Appearance

Reducing the dimensionality of material representations while capturing the statistical structure of their appearance (either across images or BSDF tabulated data) is a challenging task with multiple applications for computer graphics, such as database searching and visualization [LMS*19], gamut mapping [SSGM17] or material editing [SCW*21, DLC*22, SL23].

Some early works show that linear reduction methods such as Principal Component Analysis (PCA) are good for finding directions that capture some traits of material appearance [MPBM03, NJR15, SGM*16]. Since the emergence of deep learning, supervised learning methods have demonstrated good performance for efficient compression of complex material representations, such as Bidirectional Texture Functions (BTFs) [RJGW19, RGJW20] or tabulated BSDFs [HGC*20]. Other works also show that supervised methods are efficient in compressing images into low-dimensional representations while capturing the statistical structure of materials [LMS*19, SCW*21]. However, these data-hungry algorithms require collecting large amounts of labeled data for training. On the other hand, recent works suggest that unsupervised learning methods are also capable of compacting high-dimensional BSDF tabulated data in low-dimensional latent vectors [BSP22, ZZW*21], and keeping certain structures related to high-level perceptual attributes on simple image stimuli, such as gloss [SAF21] or translucency [LSX23]. However, aligning such latent dimensions with human perception for the case of complex stimuli remains challenging.

In our work, we propose a weakly supervised encoder that compresses images in a low-dimensional latent space capturing the structure of gloss. Our weakly supervised learning approach allows us to reduce the annotation cost, by automatically labeling training images, while still supervising the learning process to simplify the task for complex stimuli.

2.3. Material Appearance Datasets

Existing databases of materials, such as MERL [MPBM03], RGL [DJ18] and UTIA [FV14] contain measured BSDFs of different material types (e.g., metals, plastics or fabrics) and are widely used for material appearance applications [SGM*16, SSGM17, SSN18]. To capture the complex interactions of materials with other factors that influence human perception of appearance, like geometry or illumination [LSGM21], previous works rely on image-based material datasets, such as CURet [DVGNK99], OpenSurfaces [BUSB13] or Lagunas et al. [LMS*19]. These large datasets are usually annotated manually through crowd-sourcing experiments, in order to obtain ground-truth labels of perceptual attributes for material appearance, such as glossiness or metallicness.

Our training process relies on the dataset from Serrano et al. [SCW*21], which includes thousands of rendered images using a variety of real-world geometries, illuminations and measured materials, manually annotated with six perceptual attributes including glossiness. We further extend this dataset with new images of analytical materials, automatically labeled with our weak labels. Although massive crowd-sourced datasets are very useful to train learning-based models, they are not ideal for a systematic evaluation of gloss prediction models, since they can be noisy and do not cover controlled variations in factors affecting gloss perception. We thus create a new, controlled test dataset for evaluation, made up of 310 reliably annotated images.

3. Datasets

We first describe the image dataset that we use to train our model (Section 3.1). Then, we describe our new controlled test dataset to systematically evaluate gloss prediction results (Section 3.2). We include our full test dataset in the supplemental material.

3.1. Training Dataset

We use a large image-based material dataset to train our models. First, we rely on the dataset from Serrano et al. [SCW*21], which includes a variety of measured materials under real-world geometries and illuminations. We use a subset of this dataset, after removing the geometries and materials used in our test dataset and all anisotropic materials, to simplify the range of material appearances, making a total of 23,616 images. Each of these images has an associated strong label of perceived glossiness, assigned manually by humans as a single rating in the 7-point Likert scale. Note that the commonly used Likert-scale ratings already capture potential non-linearities in perception. In the following, for clarity, we always refer to this subset as the *Serrano* dataset.

We then extend our training dataset with 38,250 *new* images (never annotated by humans) including new geometries, points of view, illuminations and analytical materials using Disney's *Principled BSDF* [BS12, Bur15], for which we vary the values of the roughness r and specular s parameters in the intervals $[0, 0.5]$ and $[0.1, 5]$, respectively. The images were rendered using Mitsuba 3 [JSR*22], using a global gamma-exposure tone mapping operator to avoid introducing non-uniform contrast changes following previous work [SCW*21]. We annotate this extended dataset with our automatically-computed weak labels (see Section 4.1), with the purpose of offering an effective alternative to human labeling of a full dataset, improving generalization at a very low cost. Figure 2 shows a representative subset, and additional details are included in the supplemental material.

3.2. Test Dataset

Several datasets exist that pair images of varying material appearance with ground-truth data in the form of human judgements of gloss [SCW*21, SGM*16, DLC*22, SAF21]. Most of these datasets are built with the goal of training, testing or calibrating models

Figure 2: *Example images from our training dataset. We extend the Serrano dataset [SCW*21] with a set of new images of varied geometries, illuminations and analytical materials using Disney's Principled BSDF.*

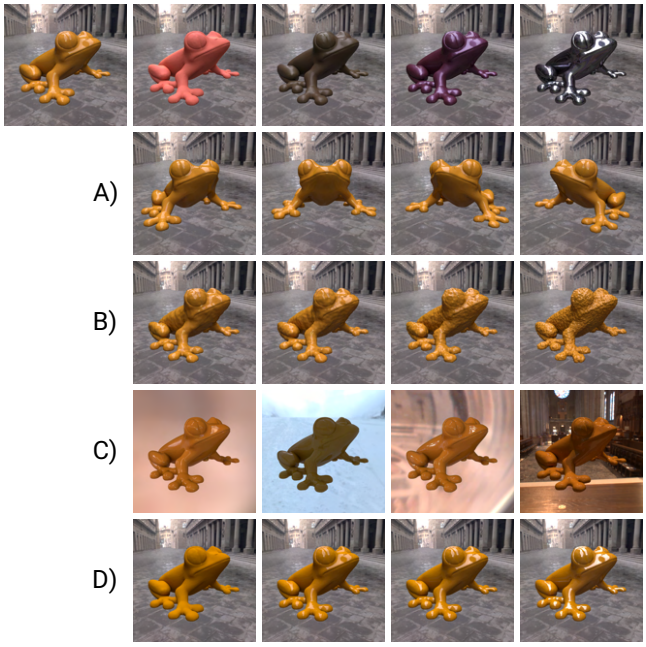

Figure 3: *Representative images of our controlled test dataset. Top row shows our five baseline measured materials under the baseline geometry* frog *and baseline illumination* uffizi. *Rows A-D show individual variations for one example material: A) different rotations, B) different geometry complexity by increasing bumpiness of the surface, C) different illuminations, and D) different levels of the specular parameter for the analytical fitting of the material.*

that mimic human perception of gloss, and feature varying material properties, lighting conditions, geometries and camera viewpoints.

However, they are not ideal for a systematic evaluation of gloss prediction models, for several reasons. First and foremost, images in the dataset do not include *controlled* variations, such as systematic changes in point of view, or illumination frequency, to measure consistency. Second, annotations are given through large crowdsourced experiments under varying, uncontrolled viewing conditions, and thus can be noisy and unreliable for specific images. And last, annotations might be highly unbalanced with respect to levels of perceived gloss, which can lead to erroneous conclusions when reporting an average error metric (e.g., a constant baseline model that always predicts the lowest gloss level for every image could get a low average error if the majority of the images in the test dataset are completely matte). Therefore, we create a *test dataset*, annotated under constant, controlled viewing conditions, which we use to reliably evaluate our models.

Our test dataset includes twenty baseline samples, from which we generate controlled variations, yielding a total of 310 images. The baseline samples are created through combinations of geometry, material (measured BSDFs), and illumination (environment maps), and cover a reasonable range of appearances (see Figure 3, top row). Specifically, our baseline includes:

- Two geometries with different complexity: *frog* and *bumpy_sphere*.
- Two illuminations with different distributions of spatial frequencies: *uffizi* and *st_peters*.
- Five measured materials, from the MERL database [MPBM03], with different glossiness: *pink_plastic*, *fruitwood*, *violet_acrylic*, *specular_yellow_phenolic* and *aluminium*.

For each combination of the baseline materials, illuminations and geometries, the following variations are generated (see Figure 3, rows A through D):

A) Five different rotations of the object.

B) Five different levels of geometry complexity, by increasing the bumpiness of the surface.
C) Three additional illuminations. Ranked in order of increasing high-frequency content [BBP04], the full set of illuminations is: *st_peters_blurred*, *glacier*, *uffizi*, *st_peters*, and *grace*. We obtain *st_peters_blurred* by applying a Gaussian filter of size 20*x*20 to the original *st_peters*, to offer a direct comparison.
D) Five different levels of the specular parameter for the Ward-Duer [Dür06] analytical fitting of the material, sampling the range [-0.05, +0.05]. Although the Ward-Duer model has limitations modeling the Fresnel effects, this model adequately fits our need to vary the specular level in the MERL database materials, as the parameters for the material fitting are publicly available in the work of Ngan et al. [NDM06].

Our resulting dataset includes challenging examples (e.g., non-conventional illumination maps like *glacier*), both measured and analytical materials, and large differences in geometry complexity. As well as providing a challenging benchmark for gloss prediction, these variations provide a means to evaluate gloss constancy, both for human annotators and for models.

Each of these 310 images is manually annotated for gloss, on a 7-point Likert scale, by five different subjects (ages 24-35, three females and two males, all claiming experience in computer graphics). Annotation is done under constant, controlled viewing conditions (SDR display and fixed lighting). Agreement between annotators is high (Krippendorff's alpha [Kri11] = 0.87; 1.0 would indicate perfect agreement), and annotations are well balanced between gloss levels (see further details in the supplemental material).

4. Gloss Prediction with Weak Labels

To obviate the need to collect a large number of strong (manual) gloss labels via time-consuming user studies, we propose combining just a small set of strong labels with a larger set of *automatically computed* weak labels.

4.1. Automatic Weak Labels

Our objective in developing automatic weak labels is to ensure their ease of computation while maintaining a rough correlation with the perception of gloss. These labels are obtained through simple and coarse approximations. To illustrate this, we explore three simple methods to automatically produce weak labels for our images, based on i) Disney's *Principled BSDF* model [BS12, Bur15]; ii) image statistics; and iii) industry metrics [HJ39], respectively. As a result, every image $\mathbf{x}$ in our analytical dataset has three different weak labels $\{y_0, y_1, y_2\}$ (in the range $[1,7]$) associated with it.

Disney's Principled BSDF Model (y_0) Previous works [PFG00, WAKB09, AKLM18] suggest that roughness r and specular s are the two main physical parameters related to gloss perception. We use a weighted combination of these parameters in Disney's Principled BSDF Model to approximate the gloss level for each image, as $y_0 = \left\lfloor \lambda_s s + \beta(\alpha - \lambda_r r^2) \right\rfloor$ where λ_s and λ_r are weights to balance the distribution of the labels, which we empirically set to 0.95 and 1.2 respectively, and β and α are scalars to translate the roughness term to the interval $[1,2]$ (set to 4 and 0.5, respectively).

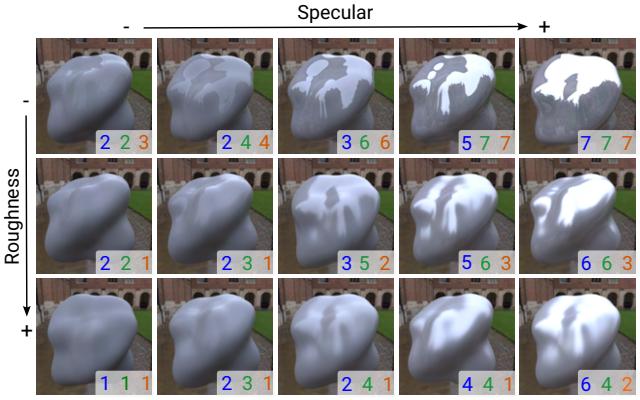

Figure 4: *Weak labels automatically computed for example images in our training dataset: based on the BSDF model (blue), image statistics (green) and industry metrics (orange). We show samples from the* blob *geometry under the* cambridge *illumination and grey albedo with different combinations of roughness and specular parameters from Disney's Principled BSDF. Although none of these labels are precise indicators of perceived gloss, they sufficiently correlate with gloss perception to be used as weak labels. All weak labels are in the range [1, 7].*

Image Statistics (y_1) Some works point to simple image statistics as indicators of human perception of gloss for simple stimuli [MNSA07, SLM*08, WTG15]. Although such approaches may not generalize well to complex images or high-frequency illumination, we analyze their usefulness as weak labels since they are easy to compute and not linked to any specific BSDF model. We define $y_1 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3/\sigma^3$ as the skewness of the luminance histogram [MNSA07], where $\bar{x}$ and σ are its mean and the standard deviation, respectively. We compute the skewness per geometry and material under low-frequency illumination in grey scale, where we empirically observe a reasonable correlation to gloss perception. To compute the skewness only from pixels inside the objects, we remove the background using a mask. Then, we assign the same label y_1 across different illuminations. Recent literature shows that more complex visual features are better correlated to human perception [SBD23] but these require access to matte and specular components separately, so we choose skewness for its simplicity.

Industry Metrics (y_2) The standard D523 from the American Society for Testing and Materials (ASTM) provides a procedure to measure gloss on real-world materials [HJ39]. Westlund and Meyer [WM01] apply its methodology to compute the ratio between the radiance given by simple BSDF models and an analytical approximation of a polished black glass. We follow a similar approach and calculate the log-ratio between the radiance computed with Disney's *Principled BSDF* (used to render our analytical dataset) and a black glass modeled by a GGX microfacet model [WMLT07], as $y_2 = \log(R_d+1)/\log(R_g+1)$, where R_d and R_g are the radiance of Disney's *Principled BSDF* and the black glass respectively, computed at an angle of 20º with respect to the normal. Both R_d and R_g are computed with the radiance meter provided by Mitsuba 3 [JSR*22]. Each resulting y_i approximates the gloss

level of **x** on a 7-point Likert scale. We show examples of our resulting weak labels in Figure 4. Although none of them is suitable to directly derive a gloss metric (nor have they been designed for that purpose), we can observe how these labels reasonably correlate with perceptual gloss, which suffices for our purposes.

4.2. Gloss Predictor

We next describe the network architecture and our weakly supervised framework for training a gloss predictor, together with implementation details for reproducibility. During training, we use data augmentation to increase the number of images by flipping, cropping, shifting, rotating, scaling, and adding Gaussian and Poisson noise.

As a feature extractor, following the work from Serrano et al. [SCW*21], we use a VGG16 architecture and remove its last layer. Then, we initialize our feature extractor with weights from pre-training on the ImageNet dataset [DDS*09] and introduce two fully connected layers to compress the features into a reduced 20-dimensional latent vector **z**. In addition, since we are interested in generating a linearly separable space, we introduce a linear regression layer that predicts the gloss level from **z**, constrained to the range $[0,1]$.

We train our gloss predictor for regression by minimizing the Mean Absolute Error (MAE) between its predicted value $\hat{y}$ and the training label y normalized (min-max normalization) to the interval $[0,1]$. This training label can be either *strong* (coming from human annotations and computed as the median between users) or *weak* (automatically computed with any of our three methods explained in Section 4.1). Therefore, our training loss L_{MAE} is defined as:

$$L_{MAE} = \frac{1}{N}\sum_{j=1}^{N} |y_j - \hat{y}_j|, \tag{1}$$

where N is the number of images in the batch. We train our predictor for 35 epochs with batch size $N = 4$ and input resolution 512x512, using the Ranger optimizer [ZLBH19, LJH*20] with an initial learning rate set to 10^{-5}.

5. Results

We evaluate our gloss predictor on our controlled test dataset described in Section 3.2. We begin by thoroughly evaluating the impact of our weakly supervised learning strategy with our three types of weak labels (Section 5.2), followed by a consistency evaluation across the controlled variations in the test dataset (Section 5.3). Then, we compare the accuracy of our gloss predictors to the current state of the art (Section 5.4) and show the ability of our models to generalize to more diverse data, including real images (Section 5.5). Finally, we perform ablation studies to validate the impact of our main design decisions (Section 5.6), and explore the structure of our 20-dimensional latent space with respect to human perception of gloss (Section 5.7).

5.1. Evaluation Metrics

To quantitatively evaluate our predictions, we follow recent related work [SCW*21] and use the Mean Absolute Error (MAE), which measures the absolute distance to the ground-truth gloss judgements of our controlled test dataset, normalized to the interval $[0,1]$. We define the ground-truth (GT) judgement per image as the median across our five human ratings, as a robust estimator of human perception. By considering the median, we obtain a more reliable estimation that is less influenced by individual variations or extreme judgments. In addition, in order to evaluate the consistency of our gloss predictions and their rank performance with respect to the ground truth, we compute the Pearson correlation, which measures the linear dependence between them without taking into account the absolute scale. Finally, to consider monotonic, but not necessarily linear, relationships we also compute Spearman's rank correlation. Both correlation coefficients are bounded to the range $[-1,1]$, where 1 means perfect positive correlation.

Table 1: *Mean Absolute Error (MAE ↓) on our controlled test dataset for our 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, S. only), combining these strong labels with our weak labels based on either BSDF model (S.+BSDF), image statistics (S.+Image stats.) or industry metrics (S.+Industry) and using only our weak labels (S. 0%) We use S. to refer to the Serrano dataset.*

	S. only	S.+BSDF	S.+Image stats.	S.+Industry
S. 100%	0.1510	**0.1207**	0.1389	0.1484
S. 20%	0.2091	0.1538	0.1550	0.1797
S. 0%	-	0.3114	0.3015	0.3466

5.2. Weakly Supervised Learning

"Same annotation cost, better performance": We show the effectiveness of our weak labels in improving gloss predictions without the need to collect a large amount of costly manual annotations. In Table 1 (first row) we show MAE results on our controlled test dataset for our gloss predictor trained only on the *Serrano* dataset with strong labels, compared to the same predictor trained on our full training dataset, including our weak labels. Although this is a challenging task, as noisy weak labels could easily degenerate learning performance even with more data [LGZ19], our weakly supervised gloss predictor outperforms the predictor trained only with strong labels, across our three different weak labels.

In Table 1 (third row) we show MAE results on our controlled test dataset featuring our gloss predictor trained exclusively using weak labels. Despite the limited ability of our weak labels *alone* to capture human gloss perception accurately, they are still sufficiently good to guide the learning process when *combined* with human-annotated data. Our weak labels allow us to train on a larger training dataset, including new geometries, illuminations, and materials, helping the model better generalize without incurring the high cost of manually annotating such new data.

"Same performance, substantially less annotation cost": Next, we study to what extent our weak labels can help mitigate the cost of collecting large human-annotated datasets, while still maintaining performance. We train our gloss predictor on a small randomly selected subset of the *Serrano* dataset, consisting of only

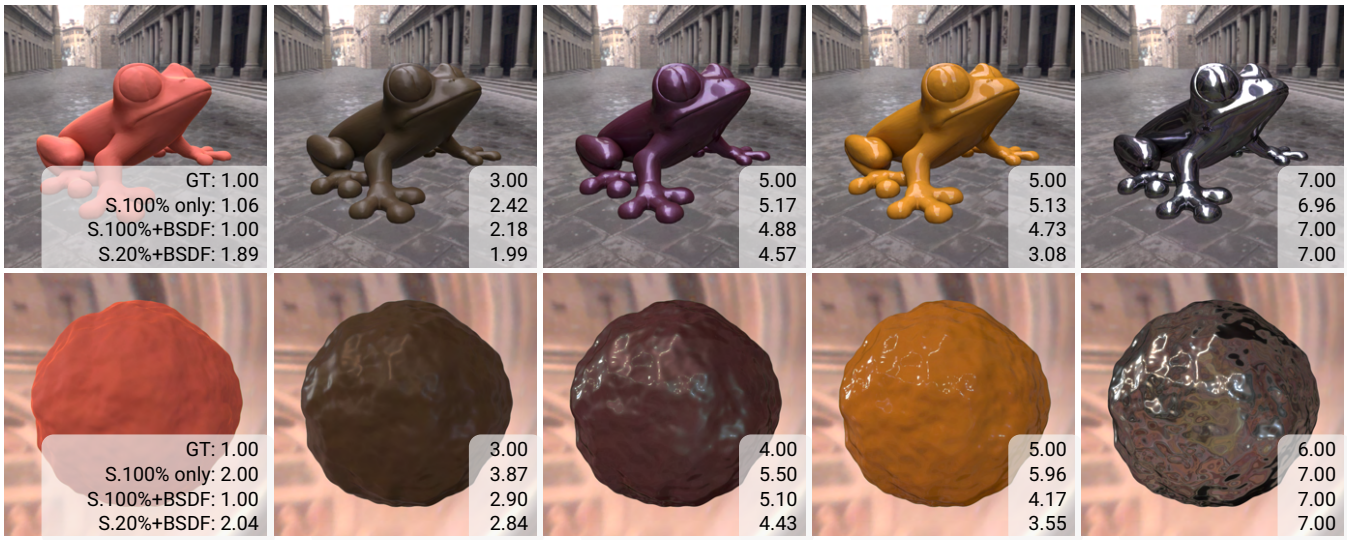

Figure 5: *Qualitative results of our gloss predictors on example images from our controlled test dataset. The numbers in the insets indicate, for every input image, the ground-truth judgement (GT) and the predictions from our models (from top to bottom): supervised model trained only with strong labels (S.100% only, second line), weakly supervised model with strong labels combined with our BSDF weak labels (S.100% + BSDF, third line) and weakly supervised model with a subset of the strong labels combined with our BSDF weak labels (S.20% + BSDF, fourth line). All gloss ratings are in the range* $[1,7]$.

20% of the strongly labeled images. We combine this reduced human-annotated subset with our extended weakly labeled dataset and show that we get similar, competitive performance (Table 1, second row). In particular, for both BSDF and image statistics weak labels we can reduce the cost of human annotation by 80%, without losing accuracy with respect to using the 100% of the *Serrano* dataset (MAE = 0.15 in S.20%+BSDF, S.20%+Image stats and S.100% only). These two different weak labels are complementary: for synthetic images with analytical materials we could rely on labels based on the BSDF model, while image statistics could be easily computed from any simple image. Hereafter, we will focus on our weakly supervised models using the BSDF weak labels (Table 1, S.100%+BSDF and S.20%+BSDF), as they are the ones that show the best performance. Refer to the supplemental material for further results using the image statistics and industry weak labels.

We show qualitative results in Figure 5. We observe how incorporating our weak labels in addition to the 100% of the strongly labeled *Serrano* dataset (S.100%+BSDF) leads to gloss predictions that better correlate with the ground-truth human judgements compared to training only on the strongly labeled data (S.100% only). Additionally, combining our weak labels with a subset of the *Serrano* dataset (S.20%+BSDF) still maintains reasonably accurate gloss predictions. More qualitative results are included in the supplemental material.

5.3. Consistency Evaluation

We evaluate the consistency of our weakly supervised gloss predictors for every variation of different confounding factors present in our test dataset, both quantitatively and qualitatively. Table 2 shows how our predictors perform consistently well across variations of rotation, bumpiness, illumination, and specularity, for the three different evaluation metrics. Although the error (MAE) is slightly higher when increasing the specular level (row D), the Pearson correlation remains high, indicating that our predictors successfully capture the trend of the perceived glossiness.

Table 2: *Quantitative results when varying each of the confounding factors in our test dataset. We include results of our weakly supervised gloss predictors, both trained with our BSDF weak labels jointly with the 100% of the Serrano dataset (S.100% + BSDF) vs. the 20% of it (S.20% + BSDF). We show MAE, Spearman and Pearson correlations results for every variation.*

Variation	MAE ↓	Spearman ↑	Pearson ↑
S.100% + BSDF			
A) Rotation	0.1043	0.9050	0.9119
B) Bumpiness	0.1076	0.8796	0.8979
C) Illumination	0.1004	0.9089	0.9129
D) Specularity	0.1448	0.8501	0.9086
S.20% + BSDF			
A) Rotation	0.1381	0.8746	0.8591
B) Bumpiness	0.1460	0.8262	0.8248
C) Illumination	0.1356	0.8565	0.8411
D) Specularity	0.1801	0.8100	0.8696

Figure 6 illustrates qualitative results for variations of rotation and specularity. The top row of images shows how human perception of gloss (GT) remains constant under different rotations of the

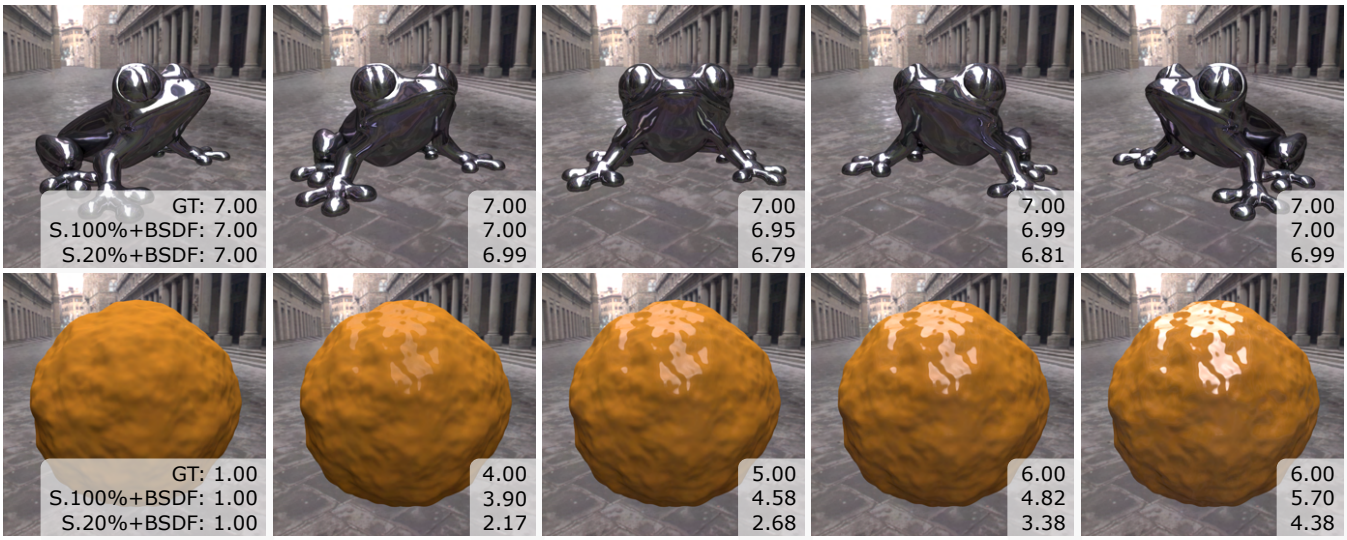

Figure 6: *Qualitative results of our weakly supervised gloss predictors when varying one confounding factor at a time on our test dataset. We show: variation across different rotations for the frog geometry with uffizi illumination and aluminium material (top), and variation across increasing specularity for the bumpy_sphere geometry with uffizi illumination and the Ward-Duer BRDF fitting of specular_yellow_phenolic material (bottom). The numbers in the insets indicate the ground-truth judgements (GT) and the gloss predictions from our weakly supervised models trained with BSDF weak labels jointly with the 100% of the Serrano dataset (S.100%+BSDF, second line) and the 20% of the Serrano dataset (20%+BSDF, third line). All gloss ratings are in the range* $[1,7]$.

object. Accurately, our weakly supervised predictions exhibit very low variation (std = 0.01 and 0.09, respectively), also preserving gloss constancy. The bottom row shows the effect of increasing the specular level; as expected, the ground-truth perceived gloss also increases (although not linearly). We observe how our predictors tend to underestimate the absolute gloss predictions (slightly for the case of the S.100%+BSDF model and more clearly for the S.20%+BSDF one), yet effectively capture the increasing trend in both cases. Further qualitative results for variations in bumpiness and illumination are included in the supplemental material.

5.4. Comparison to State of the Art

We compare our gloss predictors against a supervised state-of-the-art gloss predictor from Serrano et al. [SCW*21] (available in https://github.com/Hans1984/material-illumination-geometry) on our controlled test dataset. Results are shown in Table 3, where we see that all our predictors match significantly better with ground-truth human judgements across all evaluation metrics, even when trained on a smaller amount of human-annotated data.

5.5. Generalization

Additionally, we evaluate how well our weakly supervised predictors generalize to challenging, out of distribution data, using the test set B from Serrano et al. [SCW*21]. This test set includes 78 synthetic images depicting new materials, geometries and illuminations not included in our training data, as well as 20 "in-the-wild" real photographs of materials from the Flickr Material Database (FMD) [SRA14]. We report quantitative results in Table 4, and show illustrative examples in Figure 7. As we can see, our gloss predictors obtain a reasonable performance despite the challenging out-of-distribution dataset: mirror-like surfaces (left and middle-left) are accurately predicted by our models as highly glossy despite their complex reflections; the cloth material (right) is predicted as medium-gloss by our S.20%+BSDF model (second row), probably due to the high contrast of the knitted pattern, while our best model (S.100%+BSDF, first row) is able to successfully predict it as a low-gloss object.

Table 3: *Quantitative evaluation of our gloss predictors and Serrano et al. [SCW*21] gloss predictor on our controlled test dataset. We include results of our weakly supervised gloss predictors, both trained with our BSDF weak labels jointly with the 100% of the Serrano dataset (S.100% + BSDF) vs. the 20% of the Serrano dataset (S.20% + BSDF). We show results for MAE, Spearman and Pearson correlations with respect to the ground truth.*

Gloss predictor		MAE ↓	Spearman ↑	Pearson ↑
	Serrano et al.	0.3293	0.5662	0.5358
Ours	S.100%+BSDF	**0.1207**	**0.8594**	**0.8788**
	S.20%+BSDF	0.1538	0.8366	0.8228

5.6. Ablation Studies

We analyze how different design decisions of our training affect performance. We independently evaluate the possibility of removing our data augmentation pipeline and adding background masks

Table 4: *Quantitative evaluation of the generalization capabilities of our weakly supervised gloss predictors and Serrano et al. [SCW*21] gloss predictor on their test set B, which contains real photographs. We include results of our weakly supervised gloss predictors, both trained with our BSDF weak labels jointly with the 100% of the Serrano dataset (S.100% + BSDF) vs. the 20% of the Serrano dataset (S.20% + BSDF). We show results for MAE, Spearman and Pearson correlations with respect to the ground truth.*

Gloss predictor		MAE ↓	Spearman ↑	Pearson ↑
	Serrano et al.	0.3327	0.4546	0.4266
Ours	S.100% + BSDF	**0.2236**	**0.6625**	**0.6570**
	S.20% + BSDF	0.2386	0.6208	0.6063

Figure 7: *Qualitative results on challenging out-of-distribution images from test set B [SCW*21], that include (from left to right): a synthetic image with a very glossy mirror-like surface and a highly complex geometry, several real objects with mirror-like surface, a real image of plastic material, and a real photograph of cloth with a knitted pattern. The numbers in the insets indicate gloss predictions from our weakly supervised models trained with BSDF weak labels jointly with the 100% of the Serrano dataset (S.100%+BSDF, first row) and the 20% of the Serrano dataset (S.20%+BSDF, second row). Predictions are in the range* $[1,7]$.

to the input images (masking out the environment by setting all background pixels to 0). As shown in Table 5, the data augmentation pipeline has a major positive effect on the predictions as it helps the models to generalize better to unseen data. Background masks do not seem to help the models to estimate human gloss perception. We hypothesize that the networks might leverage the contextual cues from the background to disambiguate between material properties and illumination, similar to how humans do [AKLM18].

Additionally, we test a modified version of our training loss in Equation 1 to take into account the different classes of training labels (strong/weak); we use a loss reweighting scheme [SKP*22] and define our weighted MAE loss (L_{wMAE}) as:

$$L_{wMAE} = \frac{1}{N}\sum_{j=1}^{N} w(y_j)|y_j - \hat{y_j}|, \tag{2}$$

where $w(y_j)$ is a weighting function that depends on the class of the label y_j. Therefore, our original L_{MAE} loss is equivalent to L_{wMAE} when using a constant weighting function $w(y_j) = 1$, which does not differentiate between strong and weak labels. We show results using different weighting functions in Table 6. As expected, setting higher weights to our weakly labeled images has a negative effect; however, assigning higher weights to the strongly labeled images also deteriorates performance. We hypothesize that in these latter cases (last two rows of Table 6), the training focuses too much on

Table 5: *Ablation studies by independently changing different design elements of our models: removing data augmentation (w/o aug.) and adding background masks (bg. mask). We include results of our weakly supervised gloss predictors, both trained with our BSDF weak labels jointly with the 100% of the Serrano dataset (S.100% + BSDF) vs. the 20% of the Serrano dataset (S.20% + BSDF). We show MAE results on our test dataset.*

MAE ↓	Ours	w/o aug.	bg. mask
S.100% + BSDF	**0.1207**	0.3113	0.1819
S.20% + BSDF	**0.1538**	0.3131	0.1976

Table 6: *Ablation studies by changing the weighting function in our L_{wMAE} training loss. We include results of our weakly supervised gloss predictors, both trained with our BSDF weak labels jointly with the 100% of the Serrano dataset (S.100% + BSDF) vs. the 20% of the Serrano dataset (S.20% + BSDF). We show MAE results on our test dataset.*

$w(y_j)$		MAE ↓	
strong	weak	S.100%+BSDF	S.20%+BSDF
1.0	5.0	0.1655	0.2685
1.0	1.5	0.1634	0.1800
1.0	1.0	**0.1207**	**0.1538**
1.5	1.0	0.1361	0.1622
5.0	1.0	0.1775	0.1900

the strongly labeled images compared to the weakly labeled ones that extend the variability of the training data. Therefore, the models generalize worse, and predictions in our test set (composed of unseen scenes) become less precise. In the limit, excessively high weighting for the strongly labeled data would make it equivalent to using only the *Serrano* dataset. Further research is needed to better understand why the network overfits here to the strongly labeled data. In conclusion, our simple approach using equal weights, leads to the best results.

Finally, although our weakly-annotated images are almost free to obtain, we validate that the maintained performance is not due to the larger amount of total training images. To do so, we train several models with exactly the same amount of training images (23,616), but varying the percentage of strong labels from the *Serrano* dataset (from 0% to 100% = 23,616 strong labels). To complete the training data, we randomly select images from our extended analytical dataset labeled with our BSDF weak labels. In Figure 8, we show that 20% of strong labels (+80% of weak labels) are sufficient to achieve a similar performance as using the 100% of the *Serrano* dataset (100% of strong labels), hence showing the effectiveness of the weakly supervised strategy independent of the total number of training images.

5.7. Towards a Perceptually Meaningful Latent Space

Our gloss predictors are designed to encode input images into a low-dimensional latent space $\mathbf{z}$ of 20 dimensions. Therefore, if predictions are accurate, this latent space should capture aspects of the underlying structure of gloss perception. As we can see in Figure 9 (left), the latent space is indeed meaningfully organized with

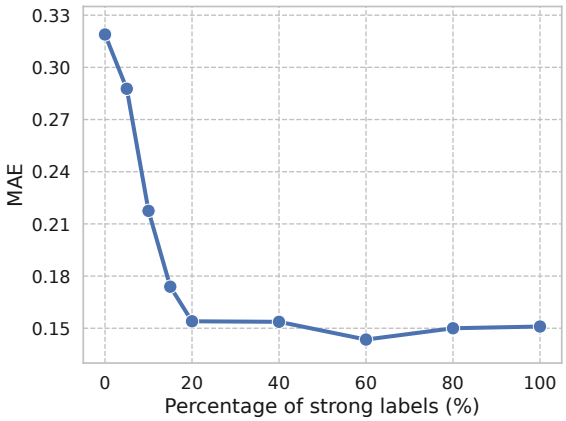

Figure 8: *Evolution of the mean absolute error (MAE ↓) when training our gloss predictor with the same amount of training images but different percentages of strong labels. 0% indicates that the predictor has been trained with 23,616 randomly selected images from our analytical dataset annotated with BSDF weak labels, while 100% indicates that the predictor has been trained with all 23,616 images from the Serrano dataset annotated with strong labels. We see that, also when the total number of training images is constant, only a 20% of strong labels is needed to achieve similar performance to using 100% of strong labels.*

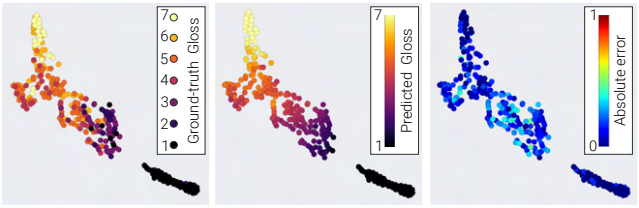

Figure 9: *Visualization of the latent space. We show all images from our controlled test dataset as points in 2D space using t-SNE [VdMH08] dimensionality reduction from the 20-dimensional feature vectors of the latent space. The color of every point indicates, from left to right: ground-truth gloss judgements, gloss predictions from our model and absolute error (normalized to the range $[0,1]$). In this figure, we use our best weakly supervised model, trained on the Serrano dataset and our BSDF weak labels (S.100%+BSDF).*

respect to human judgements of gloss for our test dataset. Very matte and very glossy materials (gloss levels 1 and 7) get clearly clustered, likely due to the higher confidence (and therefore consistency) of human labels towards the extreme values of the gloss range. The latent space is also smoothly related to the network's output predicted gloss (Figure 9, center), leading to a generally low absolute error (Figure 9, right) for all images in the test dataset. This latent space of perceptual gloss could be potentially leveraged for several applications such as material recommendations or database visualization [LMS*19].

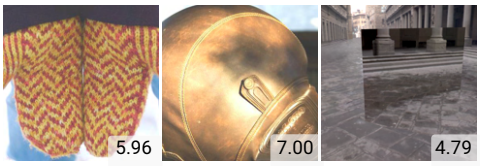

Figure 10: *Failure cases: real images with multicolored patterned surfaces (left) and very sharp shadows (center) are sometimes predicted by our model as highly glossy surfaces; perfectly-specular flat surfaces are not properly predicted by our model as mirrors (right). To illustrate these general failure cases, we use here our best weakly supervised model (S.100%+BSDF). Predictions are in the range $[1,7]$.*

6. Discussion and Future Work

This paper explores weakly supervised learning in the context of gloss prediction from images. Our work shows that, when modeling perceived material appearance, it may be possible to reduce the high cost of collecting human annotations by leveraging automatically computed weak labels for model supervision. These weak labels, combined with a reduced set of strong human labels, lead to accurate gloss predictions. Furthermore, our predictions are consistent with human perception of gloss for systematic changes in confounding factors that influence appearance, generalize reasonably well to out-of-distribution images, and exhibit an organized latent space with respect to human perception.

Our work is not free of limitations. First, although our weakly supervised gloss predictors accurately capture the trend of human ratings for systematic variations of appearance, we see how they tend to slightly underestimate the absolute gloss level, especially for images rendered with analytical materials. Additionally, despite their reasonable generalization performance for real images, our predictions fail in some challenging cases, such as multicolored patterns, very bright scenes exhibiting sharp shadows and perfectly-specular flat surfaces, as shown in Figure 10. For the first two cases, we hypothesize this might be due to the network interpreting high-contrast differences as glossy highlights. For the latter, we believe the network bases its predictions on the materials of the objects reflected on the perfectly specular surface, instead of focusing on the material of the mirror itself, which should be predicted as highly glossy. Providing an object mask as an additional channel could help the gloss predictor to divert its attention from the image background, improving its performance. However, this process can be somewhat cumbersome or imprecise, especially when working with real images.

Our work inspires promising lines of future research. First, we have shown how our simple weakly supervised strategy, using a constant weight that does not take into account the type of label, suffices for effective learning of accurate gloss predictions. It is not straightforward that weakly supervised learning could be successful in the field of gloss perception, nor what the weak labels should be. In the future, we believe that other weakly supervised learning frameworks could be exploited, including, for instance, assigning loss weights automatically based on the noise of each weak label [RZYU18, CZZ*20], or transfer learning strategies be-

tween weak and strong labels [RJS20]. We have proposed three one-dimensional weak labels, conceptually simple and inexpensive to compute. Using one-dimensional metrics is inspired by common practice in the literature that operationally defines gloss as a single judgement provided by human observers in response to "how glossy is the surface?" [SCW*21,KMA12,MKA12], that captures their overall impression. However, perceived gloss is multi-dimensional [CK15], so we believe future work could extend our approach to model multiple dimensions of gloss by defining alternative weak labels and asking observers to rate images in terms of multiple features. On the other hand, although our predictions are consistent with human perception of gloss, we evaluate our predictors on our controlled test dataset, which could be biased towards expert knowledge. Therefore, our work could be extended to evaluate whether our predictions are also consistent with the non-expert human perception of gloss, by collecting a set of test data annotated by naive observers to test whether this yields the same conclusions. In addition, our datasets (as well as the *Serrano* dataset) contain only opaque materials. We believe the current generalization limitations could be mitigated by augmenting the training data to a wider variety of optical characteristics, such as translucent or iridescent materials, as well as more diverse images (e.g., real photographs or patterned surfaces).

Finally, our weakly supervised predictors encode images into a low-dimensional latent space that is well organized with respect to perceptual gloss. Exciting future work opens up to explore how to further guide this space in order to disentangle other material appearance factors.

Acknowledgments

This work has received funding from the Spanish Agencia Estatal de Investigación (Project PID2022-141539NB-I00 funded by MCIN/AEI/10.13039/501100011033/FEDER, EU) and from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 956585 (PRIME). This research was also supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—project number 222641018—SFB/TRR 135 TP C1), by the European Research Council (ERC, Advanced Grant "STUFF"—project number ERC-2022-AdG-101098225), by the Marsden Fund of the Royal Society of New Zealand (MFP-UOA2109) and by the Cluster Project "The Adaptive Mind", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art. Julia Guerrero-Viu was supported by the FPU20/02340 predoctoral grant and J. Daniel Subias was supported by the CUS/702/2022 predoctoral grant. We also thank Daniel Martin for helping with figures and the members of the Graphics and Imaging Lab for insightful discussions.

References

[AKLM18] Adams W. J., Kucukoglu G., Landy M. S., Mantiuk R. K.: Naturally glossy: Gloss perception, illumination statistics, and tone mapping. *Journal of Vision 18*, 13 (2018).

[BBP04] Brossier P., Bello J. P., Plumbley M. D.: Real-time temporal segmentation of note objects in music signals. In *Proceedings of the International Computer Music Conference - ICMC* (2004).

[BRFFF16] Bearman A., Russakovsky O., Ferrari V., Fei-Fei L.: What's the point: Semantic segmentation with point supervision. In *Proceedings of the European Conference on Computer Vision - ECCV* (2016).

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

[BSP22] Benamira A., Shah S., Pattanaik S.: Interpretable disentangled parametrization of measured brdf with β-VAE. *arXiv preprint arXiv:2208.03914* (2022).

[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).

[BUSB13] Bell S., Upchurch P., Snavely N., Bala K.: Opensurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on Graphics 32*, 4 (2013), 1–17.

[CFJ*20] Chen Z., Fu Z., Jiang R., Chen Y., Hua X.-S.: Slv: Spatial likelihood voting for weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition - CVPR* (2020).

[CK15] Chadwick A., Kentridge R.: The perception of gloss: A review. *Vision Research 109* (2015), 221–235. Perception of Material Properties (Part I).

[CPO*17] Cho D., Park J., Oh T.-H., Tai Y.-W., So Kweon I.: Weakly-and self-supervised learning for content-aware deep image retargeting. In *Proceedings of the International Conference on Computer Vision - ICCV* (2017), pp. 4558–4567.

[CPW*22] Chen B., Piovarči M., Wang C., Seidel H.-P., Didyk P., Myszkowski K., Serrano A.: Gloss management for consistent reproduction of real and virtual objects. In *SIGGRAPH Asia 2022 Conference Papers* (2022), pp. 1–9.

[CZZ*20] Cheng L., Zhou X., Zhao L., Li D., Shang H., Zheng Y., Pan P., Xu Y.: Weakly supervised learning with side information for noisy labeled images. In *European Conference on Computer Vision - ECCV* (2020), pp. 306–321.

[DDS*09] Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L.: Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR* (2009), pp. 248–255.

[DJ18] Dupuy J., Jakob W.: An adaptive parameterization for efficient material acquisition and rendering. *Transactions on Graphics (Proceedings of SIGGRAPH Asia) 37*, 6 (2018), 274:1–274:18.

[DLC*22] Delanoy J., Lagunas M., Condor J., Gutierrez D., Masia B.: A generative framework for image-based editing of material appearance using perceptual attributes. In *Computer Graphics Forum* (2022), vol. 41 (1), pp. 453–464.

[DLG*20] Delanoy J., Lagunas M., Galve I., Gutierrez D., Serrano A., Fleming R., Masia B.: The role of objective and subjective measures in material similarity learning. In *ACM SIGGRAPH 2020 Posters* (2020).

[Dür06] Dür A.: An improved normalization for the ward reflectance model. *Journal of Graphics Tools 11*, 1 (2006), 51–59.

[DVGNK99] Dana K. J., Van Ginneken B., Nayar S. K., Koenderink J. J.: Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics 18*, 1 (1999), 1–34.

[FDA01] Fleming R. W., Dror R. O., Adelson E. H.: How do humans determine reflectance properties under unknown illumination? *In Proceedings of IEEE Workshop on Identifying Objects Across Variations in Lighting, Lihue, HI.* (2001).

[FDA03] Fleming R. W., Dror R. O., Adelson E. H.: Real-world illumination and the perception of surface reflectance properties. *Journal of Visual Communication and Image Representation 3*, 5 (2003).

[Fle14] Fleming R. W.: Visual perception of materials and their properties. *Vision research 94* (2014), 62–75.

[Fle17] FLEMING R. W.: Material perception. *Annual review of vision science 3* (2017), 365–388.

[FS19] FLEMING R. W., STORRS K. R.: Learning to see stuff. *Current Opinion in Behavioral Sciences 30* (2019), 100–108. Visual perception.

[FV14] FILIP J., VÁVRA R.: Template-based sampling of anisotropic brdfs. In *Computer Graphics Forum* (2014), vol. 33 (7), pp. 91–99.

[H*37] HUNTER R. S., ET AL.: Methods of determining gloss. *NBS Research paper RP 958* (1937).

[HGC*20] HU B., GUO J., CHEN Y., LI M., GUO Y.: Deepbrdf: A deep representation for manipulating measured brdf. In *Computer Graphics Forum* (2020), vol. 39 (2), pp. 157–166.

[HJ39] HUNTER R. S., JUDD D. B.: Development of a method of classifying paints according to gloss. *ASTM Bulletin*, 97 (1939), 11–18.

[JSR*22] JAKOB W., SPEIERER S., ROUSSEL N., NIMIER-DAVID M., VICINI D., ZELTNER T., NICOLET B., CRESPO M., LEROY V., ZHANG Z.: Mitsuba 3 renderer, 2022. https://mitsuba-renderer.org.

[KJHH23] KIM B., JEONG J., HAN D., HWANG S. J.: The devil is in the points: Weakly semi-supervised instance segmentation via point-guided mask representation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR* (2023), pp. 11360–11370.

[KMA12] KIM J., MARLOW P. J., ANDERSON B. L.: The dark side of gloss. *Nature neuroscience 15*, 11 (2012), 1590–1595.

[Kri11] KRIPPENDORFF K.: Computing krippendorff's alpha-reliability. *Departmental Papers (ASC) 43. University of Pennsylvania.* (2011).

[LBFS21] LAVOUÉ G., BONNEEL N., FARRUGIA J.-P., SOLER C.: Perceptual quality of brdf approximations: dataset and metrics. *Computer Graphics Forum 40*, 2 (2021), 327–338.

[LGZ19] LI Y.-F., GUO L.-Z., ZHOU Z.-H.: Towards safe weakly supervised learning. *IEEE transactions on pattern analysis and machine intelligence 43*, 1 (2019), 334–346.

[LJH*20] LIU L., JIANG H., HE P., CHEN W., LIU X., GAO J., HAN J.: On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations* (2020).

[LMS*19] LAGUNAS M., MALPICA S., SERRANO A., GARCES E., GUTIERREZ D., MASIA B.: A similarity measure for material appearance. *ACM Transactions on Graphics 38*, 4 (2019).

[LSGM21] LAGUNAS M., SERRANO A., GUTIERREZ D., MASIA B.: The joint role of geometry and illumination on material recognition. *Journal of Vision 21*, 2 (2021), 2–2.

[LSX23] LIAO C., SAWAYAMA M., XIAO B.: Unsupervised learning reveals interpretable latent representations for translucency perception. *PLOS Computational Biology 19*, 2 (2023), e1010878.

[LY20] LUO W., YANG M.: Semi-supervised semantic segmentation via strong-weak dual-branch network. In *Proceedings of the European Conference on Computer Vision - ECCV* (2020).

[MKA12] MARLOW P. J., KIM J., ANDERSON B.: The perception and misperception of specular surface reflectance. *Current Biology 22*, 20 (2012), 1909–1913.

[MNSA07] MOTOYOSHI I., NISHIDA S., SHARAN L., ADELSON E. H.: Image statistics and the perception of surface qualities. *Nature 447*, 7141 (2007), 206–209.

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

[NDM06] NGAN A., DURAND F., MATUSIK W.: Image-driven Navigation of Analytical BRDF Models. In *Eurographics Symposium on Rendering* (2006).

[NJR15] NIELSEN J. B., JENSEN H. W., RAMAMOORTHI R.: On optimal, minimal brdf sampling for reflectance acquisition. *ACM Transactions on Graphics 34*, 6 (2015).

[PFG00] PELLACINI F., FERWERDA J. A., GREENBERG D. P.: Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (2000), p. 55–64.

[RGJW20] RAINER G., GHOSH A., JAKOB W., WEYRICH T.: Unified neural encoding of btfs. In *Computer Graphics Forum* (2020), vol. 39 (2), pp. 167–178.

[RJGW19] RAINER G., JAKOB W., GHOSH A., WEYRICH T.: Neural btf compression and interpolation. In *Computer Graphics Forum* (2019), vol. 38 (2), pp. 235–244.

[RJS20] ROBINSON J., JEGELKA S., SRA S.: Strength from weakness: Fast learning using weak supervision. In *International Conference on Machine Learning - ICML* (2020), pp. 8127–8136.

[RYY*20] REN Z., YU Z., YANG X., LIU M.-Y., LEE Y. J., SCHWING A. G., KAUTZ J.: Instance-aware, context-focused, and memory-efficient weakly supervised object detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR* (2020), pp. 10598–10607.

[RZYU18] REN M., ZENG W., YANG B., URTASUN R.: Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning - ICML* (2018), pp. 4334–4343.

[SAF21] STORRS K. R., ANDERSON B. L., FLEMING R. W.: Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour 5*, 10 (2021), 1402–1417.

[SBD23] SCHMID A. C., BARLA P., DOERSCHNER K.: Material category of visual objects computed from specular image structure. *Nature Human Behaviour 7*, 7 (2023), 1152–1169.

[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.

[SGM*16] SERRANO A., GUTIERREZ D., MYSZKOWSKI K., SEIDEL H.-P., MASIA B.: An intuitive control space for material appearance. *ACM Transactions on Graphics 35*, 6 (2016).

[SKP*22] SONG H., KIM M., PARK D., SHIN Y., LEE J.-G.: Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[SL23] SUBIAS J. D., LAGUNAS M.: In-the-wild Material Appearance Editing using Perceptual Attributes. *Computer Graphics Forum* (2023).

[SLM*08] SHARAN L., LI Y., MOTOYOSHI I., NISHIDA S., ADELSON E. H.: Image statistics for surface reflectance perception. *J. Opt. Soc. Am. A 25*, 4 (2008), 846–865.

[SRA14] SHARAN L., ROSENHOLTZ R., ADELSON E. H.: Accuracy and speed of material categorization in real-world images. *Journal of Vision 14*, 10 (2014).

[SSGM17] SUN T., SERRANO A., GUTIERREZ D., MASIA B.: Attribute-preserving gamut mapping of measured brdfs. In *Computer Graphics Forum* (2017), vol. 36 (4), pp. 47–54.

[SSN18] SOLER C., SUBR K., NOWROUZEZAHRAI D.: A versatile parameterization for measured material manifolds. In *Computer Graphics Forum* (2018), vol. 37 (2), pp. 135–144.

[VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-sne. *Journal of machine learning research 9*, 11 (2008).

[WAKB09] WILLS J., AGARWAL S., KRIEGMAN D., BELONGIE S.: Toward a perceptual space for gloss. *ACM Transactions on Graphics 28*, 4 (2009).

[WM01] WESTLUND H. B., MEYER G. W.: Applying appearance standards to light reflection models. In *Proceedings of Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)* (2001).

[WMLT07] WALTER B., MARSCHNER S. R., LI H., TORRANCE K. E.: Microfacet Models for Refraction through Rough Surfaces. In *Rendering Techniques - The Eurographics Association* (2007).

[WTG15] WIEBEL C. B., TOSCANI M., GEGENFURTNER K. R.: Statistical correlates of perceived gloss in natural images. *VR 115* (2015), 175–187. Perception of Material Properties (Part II).

[XXY*15] XIAO T., XIA T., YANG Y., HUANG C., WANG X.: Learning from massive noisy labeled data for image classification. In *Proceedings of the Conference on Computer Vision and Pattern Recognition - CVPR* (2015).

[Zho18] ZHOU Z.-H.: A brief introduction to weakly supervised learning. *National Science Review 5*, 1 (2018), 44–53.

[ZLBH19] ZHANG M., LUCAS J., BA J., HINTON G. E.: Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems 32* (2019).

[ZZW*21] ZHENG C., ZHENG R., WANG R., ZHAO S., BAO H.: A Compact Representation of Measured BRDFs Using Neural Processes. *ACM Transactions on Graphics 41*, 2 (2021), 1–15.