A Similarity Measure for Material Appearance

MANUEL LAGUNAS, Universidad de Zaragoza, I3A, Spain
SANDRA MALPICA, Universidad de Zaragoza, I3A, Spain
ANA SERRANO, Universidad de Zaragoza, I3A, Spain
ELENA GARCES, Universidad Rey Juan Carlos, Spain
DIEGO GUTIERREZ, Universidad de Zaragoza, I3A, Spain
BELEN MASIA, Universidad de Zaragoza, I3A, Spain

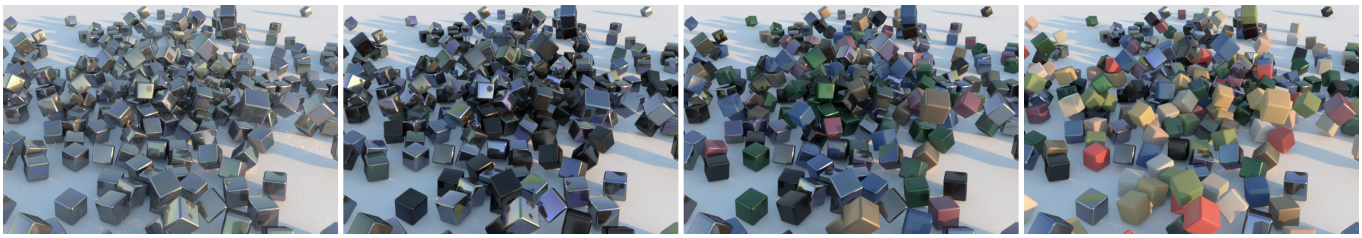

Fig. 1. The cubes in the leftmost image have all been rendered with the same aluminium material. Our similarity measure for material appearance can be used to automatically generate alternative depictions of the same scene, where the similarity of the materials varies in a controlled manner. The next three images show results with materials randomly chosen by progressively extending the search distance from the original aluminium, from similar in appearance to farther away materials within the same dataset.

We present a model to measure the similarity in appearance between different materials, which correlates with human similarity judgments. We first create a database of 9,000 rendered images depicting objects with varying materials, shape and illumination. We then gather data on perceived similarity from crowdsourced experiments; our analysis of over 114,840 answers suggests that indeed a shared perception of appearance similarity exists. We feed this data to a deep learning architecture with a novel loss function, which learns a feature space for materials that correlates with such perceived appearance similarity. Our evaluation shows that our model outperforms existing metrics. Last, we demonstrate several applications enabled by our metric, including appearance-based search for material suggestions, database visualization, clustering and summarization, and gamut mapping.

CCS Concepts: • **Computing methodologies** → *Appearance and texture representations*; *Perception*; • **Computer systems organization** → *Neural networks*.

Additional Key Words and Phrases: Material appearance, neural networks, physically based material perception

Authors' addresses: Manuel Lagunas, Universidad de Zaragoza, I3A, Spain, Zaragoza, mlagunas@unizar.es; Sandra Malpica, Universidad de Zaragoza, I3A, Spain, Zaragoza, smalpica@unizar.es; Ana Serrano, Universidad de Zaragoza, I3A, Spain, Zaragoza, anase@unizar.es; Elena Garces, Universidad Rey Juan Carlos, Spain, Madrid, elena.garces@urjc.es; Diego Gutierrez, Universidad de Zaragoza, I3A, Spain, Zaragoza, diegog@unizar.es; Belen Masia, Universidad de Zaragoza, I3A, Spain, Zaragoza, bmasia@unizar.es.

0730-0301/2019/7-ART135 $15.00
https://doi.org/10.1145/3306346.3323036

ACM Reference Format:
Manuel Lagunas, Sandra Malpica, Ana Serrano, Elena Garces, Diego Gutierrez, and Belen Masia. 2019. A Similarity Measure for Material Appearance. *ACM Trans. Graph.* 38, 4, Article 135 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323036

1 INTRODUCTION

Humans are able to recognize materials, compare their appearance, or even infer many of their key properties effortlessly, just by briefly looking at them. Many works propose classification techniques, although it seems clear that labels do not suffice to capture the richness of our subjective experience with real-world materials [Fleming 2017]. Unfortunately, the underlying perceptual process of material recognition is complex, involving many distinct variables; such process is not yet completely understood [Anderson 2011; Fleming 2014; Maloney and Brainard 2010].

Given the large number of parameters involved in our perception of materials, many works have focused on individual attributes (such as the perception of gloss [Pellacini et al. 2000; Wills et al. 2009], or translucency [Gkioulekas et al. 2015]), while others have focused on particular applications like material synthesis [Zsolnai-Fehér et al. 2018], editing [Serrano et al. 2016], or filtering [Jarabo et al. 2014]. However, the fundamentally difficult problem of establishing a *similarity measure for material appearance* remains an open problem. Material appearance can be defined as "the visual impression we have of a material" [Dorsey et al. 2010]; as such, it depends not only on the BRDF of the material, but also on external factors like lighting or geometry, as well as human judgement [Adelson 2001; Fleming 2014]. This is different from the common notion of image similarity (devoted to finding detectable differences between images, e.g., [Wang et al. 2004]), or from similarity in BRDF space (which has

been shown to correlate poorly with human perception, e.g., [Serrano et al. 2016]). Given the ubiquitous nature of photorealistic computer-generated imagery, and emerging fields like computational materials, a similarity measure of material appearance could be valuable for many applications.

Capturing a human notion of perceptual similarity in different contexts has been an active area of research recently [Agarwal et al. 2007; Garces et al. 2014; Lun et al. 2015]. In this paper we develop a novel image-based material appearance similarity measure derived from a learned feature space. This is challenging, given the subjective nature of the task, and the interplay of confounding factors like geometry or illumination in the final perception of appearance. Very recent work suggests that perceptual similarity may be an emergent property, and that deep learning features can be trained to learn a representation of the world that correlates with perceptual judgements [Zhang et al. 2018]. Inspired by this, we rely on a combination of large amounts of images, crowdsourced data, and deep learning. In particular, we create a diverse collection of 9,000 stimuli using the measured, real-world materials in the MERL dataset [Matusik et al. 2003], which covers a wide variety of isotropic appearances, and a combination of different shapes and environment maps. Using triplets of images, we gather information through Mechanical Turk, where participants are asked which of two given examples has a more similar appearance to a reference. Given our large stimuli space, we employ an adaptive sampling scheme to keep the number of triplets manageable. From this information, we learn a model of material appearance similarity using a combined loss function that enforces learning of the appearance similarity information collected from humans, and the main features that describe a material in an image; this allows us to learn the notion of material appearance similarity explained above, dependent on both the visual impression of the material, and the actual physical properties of it.

To evaluate our model, we first confirm that humans do provide reliable answers, suggesting a shared perception of material appearance similarity, and further motivating our similarity measure. We then compare the performance of our model against humans: Despite the difficulty of our goal, our model performs on par with human judgements, yielding results better aligned with human perception than current metrics. Last, we demonstrate several applications that directly benefit from our metric, such as material suggestions, database visualization, clustering and summarization, or gamut mapping. In addition to the 9,000 rendered images, our database also includes surface normals, depth, transparency, and ambient occlusion maps for each one, while our collected data contains 114,840 answers; we provide both, along with our pre-trained deep learning framework, in order to help future studies on the perception of material appearance[1].

2 RELATED WORK

2.1 Material perception

There have been many works aiming to understand the perceptual properties of BRDFs [Anderson 2011; Fleming 2014; Fleming et al. 2015; Maloney and Brainard 2010]; a comprehensive review can be found in the work of Thompson and colleagues [2011]. Finding a direct mapping between perceptual estimates and the physical material parameters is a hard task involving many dimensions, not necessarily correlated. Many researchers focus on the perception of one particular property of a given material (such as glossiness [Chadwick and Kentridge 2015; Pellacini et al. 2000; Wills et al. 2009], translucency [Gkioulekas et al. 2015, 2013], or viscosity [Van Assen et al. 2018]), or one particular application (such as filtering [Jarabo et al. 2014], computational aesthetics [Cunningham et al. 2007], or editing [Mylo et al. 2017; Serrano et al. 2016]). Leung and Malik [2001] study the appearance of flat surfaces based on textural information. Other recent works analyze the influence on material perception of external factors such as illumination [Ho et al. 2006; Křivánek et al. 2010; Vangorp et al. 2017], motion [Doerschner et al. 2011], or shape [Havran et al. 2016; Vangorp et al. 2007].

A large body of work has been devoted to analyzing the relationships between different materials, and deriving low-dimensional perceptual embeddings [Matusik et al. 2003; Serrano et al. 2016; Soler et al. 2018; Wills et al. 2009]. These embeddings can be used to derive material similarity metrics, which are useful to determine if two materials convey the same appearance, and thus benefit a large number of applications (such as BRDF compression, fitting, or gamut mapping). A number of works have proposed different metrics, computed either directly over measured BRDFs [Fores et al. 2012; Ngan et al. 2005], in image space [Ngan et al. 2006; Pereira and Rusinkiewicz 2012; Sun et al. 2017], or in reparametrizations of BRDF spaces based on perceptual traits [Pellacini et al. 2000; Serrano et al. 2016]. Our work is closer to the latter; however, rather than analyzing perceptual traits in isolation, we focus on the overall appearance of materials, and derive a similarity measure that correlates with the notion of material similarity as perceived by humans.

2.2 Learning to recognize materials

Image patches have been shown to contain enough information for material recognition [Schwartz and Nishino 2018], and several works have leveraged this to derive learning techniques for material recognition tasks. Bell et al. [2015] introduce a CNN-based approach for local material recognition using a large annotated database, while Schwartz and Nishino explicitly introduce global contextual cues [2016]. Other works add more information such as known illumination, depth, or motion. Georgoulis et al. [2017] use both an object's image and its geometry to create a full reflectance map, which is later used as an input to a four-class coarse classifier (metal, paint, plastic or fabric). For a comprehensive study on early material recognition systems and latest advances, we refer to the reader to the work of Fleming [2017]. These previous works focus mainly on classification tasks, however *mere labels do not capture the richness of our subjective experience of materials in the real world* [Fleming 2017].

Recent work has shown the extraordinary ability of deep features to match human perception in the assessment of perceptual similarity between two images [Zhang et al. 2018]. Together with the success of the works mentioned above, this provides motivation for the combination of user data and deep learning that we propose in this work.

[1] http://webdiis.unizar.es/~mlagunas/publication/material-similarity/

2.3 Existing datasets

Early image-based material datasets include CURet [Dana et al. 1999], KTH-TIPS [Hayman et al. 2004], or FMD [Sharan et al. 2009]. OpenSurfaces [Bell et al. 2013] includes over 20,000 real-world images, with surface properties annotated via crowdsourcing. This dataset has served as a baseline to others, such as the Materials in Context Database (MINC) [Bell et al. 2015], an order of magnitude larger; SynBRDF [Kim et al. 2017], with 5,000 rendered materials randomly sampled from OpenSurfaces; or MaxBRDF dataset [Vidaurre et al. 2019], which includes synthetic anisotropic materials.

Databases with *measured* materials include MERL [Matusik et al. 2003] for isotropic materials, UTIA [Filip and Vávra 2014] for anisotropic ones, the Objects under Natural Illumination Database [Lombardi and Nishino 2012], which includes calibrated HDR information, or the recent, on-going database by Dupuy and Jakob which measures spectral reflectance [2018]. We choose as a starting point the MERL dataset, since it contains a wider variety of isotropic materials, and it is still being successfully used in many applications such as gamut mapping [Sun et al. 2017], material editing [Serrano et al. 2016; Sun et al. 2018], BRDF parameterization [Soler et al. 2018], or photometric light source estimation [Lu et al. 2018].

3 MATERIALS DATASET

3.1 Why a new materials dataset?

To obtain a meaningful similarity measure of material appearance we require a large dataset with the following characteristics:

- Data for a wide variety of materials, shapes, and illumination conditions.
- Samples featuring the *same* material rendered under different illuminations and with different shapes.
- Materials represented by measured BRDFs, with reflectance data captured from real materials.
- Real-world illumination, as given by captured environment maps.
- A large number of samples, amenable to learning-based frameworks.

These characteristics enable renditions of complex, realistic appearances and will be leveraged to train our model, explained in Section 5. To our knowledge, none of the existing material datasets features all these characteristics.

3.2 Description of the dataset

In the following, we briefly describe the characteristics of our dataset, and refer the reader to the supplementary material for further details.

Materials. Our dataset includes all 100 materials from the MERL BRDF database [Matusik et al. 2003]. This database was chosen as starting point since it includes real-world, measured reflectance functions covering a wide range of reflectance properties and types of materials, including paints, metals, fabrics, or organic materials, among others.

Illumination. We use six natural illumination environments, since they are favored by humans in material perception tasks [Fleming et al. 2003]. They include a variety of scenes, ranging from indoor scenarios to urban or natural landscapes, as high-quality HDR environment maps[2].

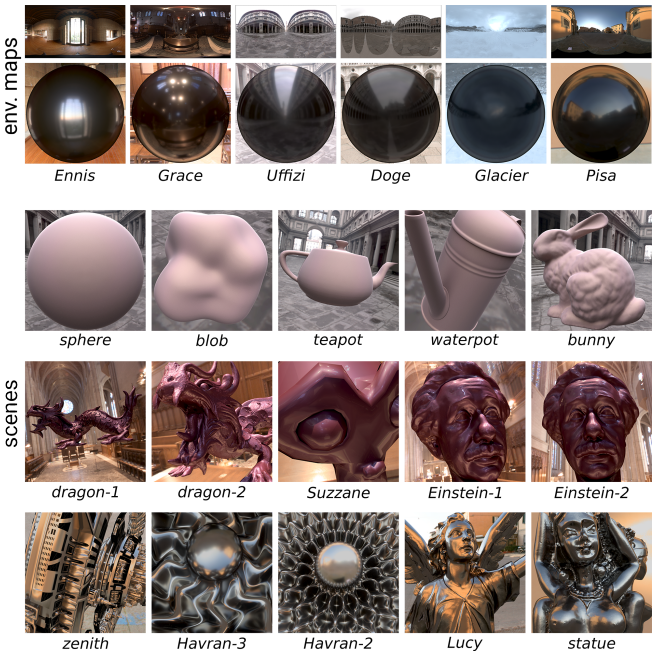

Fig. 2. **Top**: The six environment maps used in the dataset, and corresponding rendered spheres with the *black-phenolic* material. **Bottom**: Sample images of all 15 scenes with different materials and illumination conditions. First row: *pink-felt* and *Uffizi*; second row: *violet-acrylic* and *Grace*; third row: *nickel* and *Pisa*. The 3D models *bunny*, *dragon*, *Lucy* and *statue* belong to The Stanford 3D Scanning Repository; *waterpot* (modelled by gykservy), *Suzzane* (killzone75), *Einstein* (oliverlaric), and *zenith* (KuhnIndustries) were obtained from TurboSquid.

Scenes. Our database contains thirteen different 3D models, with an additional camera viewpoint for two of them, defining our fifteen scenes. It includes widely used 3D models, and objects that have been specifically designed for material perception studies [Havran et al. 2016; Vangorp et al. 2007]. The viewpoints have been chosen to cover a wide range of features such as varying complexity, convexity, curvature, and coverage of incoming and outgoing light directions.

By combining the aforementioned materials (100), illumination conditions (6), and scenes (15), we generate a total of 9,000 dataset samples using the Mitsuba physically-based renderer [Jakob 2010]. For each one we provide: The rendered HDR image, a corresponding LDR image[3], along with depth, surface normals, alpha channel, and ambient occlusion maps. While not all these maps are used in the present work, we make them available with the dataset should they be useful for future research. Figure 2 shows sample images for all fifteen scenes.

[2] http://gl.ict.usc.edu/Data/HighResProbes/

[3] Tone-mapped using the algorithm by Mantiuk et al. [2008], with the predefined *lcd office* display, and color saturation and contrast enhancement set to 1.

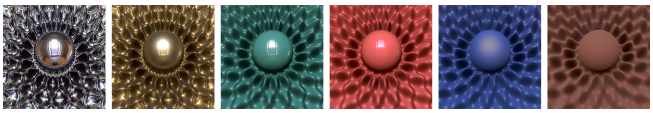

Fig. 3. Sample stimuli for our appearance similarity collection. They correspond to the *Havran-2* scene, with materials from the MERL database, rendered with the *Ennis* environment map. In reading order: *chrome*, *gold-metallic-paint3*, *specular-green-phenolic*, *maroon-plastic*, *dark-blue-paint* and *light-brown-fabric*.

4 COLLECTING APPEARANCE SIMILARITY INFORMATION

We describe here our methodology to gather crowdsourced information about the perception of material appearance.

Stimuli. We use 100 different stimuli, covering all 100 materials in the dataset, rendered with the *Ennis* environment map. We choose the *Havran-2* scene, since its shape has been designed to maximize the information relevant for material appearance judgements by optimizing the coverage of incoming and outgoing light directions sampled [Havran et al. 2016]. Figure 3 shows some examples.

Participants. A total of 603 participants took part in the test through the Mechanical Turk (MTurk) platform, with an average age of 32, and 46.27% female. Users were not aware of the purpose of the experiment.

Procedure. Our study deals with the *perception* of material appearance, which may not be possible to represent in a linear scale; this advises against ranking methods [Kendall and Babington-Smith 1940]. We thus gather data in the form of relative comparisons, following a 2AFC scheme; the subject is presented with a triplet made up of one *reference* material, and two *candidate* materials, and their task is to answer the question *Which of these two candidates has a more similar appearance to the reference?* by choosing one among the two candidates. This approach has several additional advantages: it is easier for humans than providing numerical distances [McFee and Lanckriet 2011; Schultz and Joachims 2003], while it reduces fatigue and avoids the need to reconcile different scales of similarity among subjects [Kendall and Gibbons 1990].

However, given our 100 different stimuli, a naive 2AFC test would require 495,000 comparisons, which is intractable even if not all subjects see all comparisons. To ensure robust statistics, we aim to obtain five answers for each comparison, which would mean testing a total of 2,475,000 comparisons. Instead, we use an iterative *adaptive sampling* scheme [Tamuz et al. 2011]: For any given reference, each following triplet is chosen to maximize the information gain, given the preceding responses (please refer to the supplementary material for a more detailed description of the method). From an initial random sampling, we perform 25 iterations as recommended by Tamuz et al. for datasets our size; in each iteration we sample 10 new pairs for every one of our 100 reference materials, creating 1,000 new triplets. After this process, the mean information gain per iteration is less than 10^{-5}, confirming the convergence of the sampling scheme. This scheme allows us to drastically reduce the number of required comparisons, while providing a good approximation to sampling the full set of triplets.

Each test (HIT in MTurk terminology) consisted of 110 triplets. To minimize worker unreliability [Welinder et al. 2010], each HIT was preceded by a short training session that included a few trial comparisons with obvious answers [Garces et al. 2014; Rubinstein et al. 2010]. In addition, ten control triplets were included in each HIT, testing repeated-trial consistency within participants. We adopt a conservative approach and reject participants with two or more different answers. In the end, we obtained 114,840 valid answers, yielding a participants' consistency of 84.7%.

As a separate test, to validate how well our collected answers generalize to other shapes and illuminations, we repeated the same comparisons, this time with symmetric and asymmetric triplets chosen randomly from our dataset, with the condition that they do not contain the *Havran-2* shape nor the *Ennis* illumination. For symmetric triplets, the three items in the triplet differ only in the material properties, while in asymmetric triplets geometry and lighting also vary. We launched 2,500 symmetric triplets, and found that participants' majority matched the previous responses with a 84.59% rate. When we added the same number of asymmetric triplets to the test, participants' answers held with a 80% match rate.

5 LEARNING PERCEIVED SIMILARITY

This section describes our approach to learn perceived similarity for material appearance. Given an input image ψ, our model provides a feature vector $f(\psi)$ that transforms the input image into a feature space well aligned with human perception.

We use the ResNet architecture [He et al. 2016], based on its generalization capabilities and its proven performance on image-related tasks. The novelty of this architecture is a residual block meant for learning a residual mapping between the layers, instead of a direct mapping, which enables training very deep networks (hundreds of layers) with outstanding performance. For training we use image data from our materials dataset (Section 3), together with human data on perceived similarity (Section 4). We first describe our combined loss function, then our training procedure.

5.1 Loss function

We train our model using a loss function consisting of two terms, equally weighted:

$$\mathcal{L} = \mathcal{L}_{TL} + \mathcal{L}_{P} \tag{1}$$

The two terms represent a perceptual triplet loss, and a similarity term, respectively. The terms aim at learning appearance similarity from the participants' answers, while extracting the main features defining the material depicted in an image. In the following, we describe these terms and their contribution.

5.1.1 Triplet loss term $\mathcal{L}_{TL}$. This term allows to introduce the collected MTurk information on appearance similarity. Let $\mathcal{A} = \{(r_i, a_i, b_i)\}$ be the set of answered relative comparisons, where r is the reference image, a is the candidate image chosen by the majority of users as being more similar to r, and b the other candidate; i indexes over all the relative comparisons. Intuitively, r and a should be closer together in the learned feature space than r and b. It is not feasible to collect user answers for all possible comparisons (n

different images would lead to $n\binom{n-1}{2}$ tests); however, as we have shown in Section 4, the collected answers for a triplet (r, a, b) involving materials m^r, m^a and m^b generalize well to other combinations of shape and illumination from our dataset involving the same set of materials. We thus define $\mathcal{A}^M = \{(m_i^r, m_i^a, m_i^b)\}$ as the set of relative comparisons with collected answers (m^a represents the material chosen by the majority of participants). We then formulate the first term as a triplet loss [Cheng et al. 2016; Lagunas et al. 2018; Schroff et al. 2015]:

$$\mathcal{L}_{TL} = \frac{1}{|\mathcal{B}^A|} \sum_{(r,a,b)\in\mathcal{B}^A} \left[||f(r)-f(a)||_2^2 - ||f(r)-f(b)||_2^2 + \mu\right]_+ \quad (2)$$

where $f(\psi)$ is the feature vector of image ψ, and the set $\mathcal{B}^A$ is defined as:

$$\mathcal{B}^A = \left[(r, a, b) \mid (m^r, m^a, m^b) \in \mathcal{A}^M \ \wedge \ (r, a, b) \in \mathcal{B}\right] \quad (3)$$

with $\mathcal{B}$ the current training batch. In Eq. 2, μ represents the margin, which accounts for how much we aim to separate the samples in the feature space.

5.1.2 Similarity term $\mathcal{L}_P$. We introduce a second loss term that maximizes the log-likelihood of the model choosing the same material as humans. We define this probability p_{ra} (and conversely p_{rb}) as a quotient between similarity values s_{ra} and s_{rb}:

$$p_{ra} = \frac{s_{ra}}{s_{rb} + s_{ra}}, \quad p_{rb} = \frac{s_{rb}}{s_{rb} + s_{ra}}$$

These similarities are derived from the distances between r, a and b in the feature space, where a similarity value of 1 means perfect similarity and a value of 0 accounts for total dissimilarity:

$$s_{ra} = \frac{1}{1 + d_{ra}}, \quad s_{rb} = \frac{1}{1 + d_{rb}}, \quad \text{where}$$

$$d_{ra} = ||f(r) - f(a)||_2^2, \quad d_{rb} = ||f(r) - f(b)||_2^2$$

With this, we can formulate the similarity term as:

$$\mathcal{L}_P = -\frac{1}{|\mathcal{B}^A|} \sum_{(r,a,b)\in\mathcal{B}^A} \log p_{ra} \quad (4)$$

5.2 Training details

For training, we remove the *Havran-2* and *Havran-3* scenes from the dataset, leading to 7,800 images (13 (scenes) × 6 (env. maps) × 100 (materials)), augmented to 39,000 using crops, flips, and rotations. These 39,000 images, together with the collected MTurk answers, constitute our training data. We use the corrected *Adam* optimization [Kingma and Ba 2014; Reddi et al. 2019] with a learning rate that starts at 10^{-3} to train the network. We train for 80 epochs and the learning rate is reduced by a factor of 10 every 20 epochs. For initialization, we use the weights of the pre-trained model [He et al. 2016] on ImageNet [Deng et al. 2009; Russakovsky et al. 2015]. To adapt the network to our loss function, we remove the last layer of the model and introduce a fully-connected (*fc*) layer that outputs a 128-dimensional feature vector $f(\psi)$. We use a margin $\mu = 0.3$ for the triplet loss term $\mathcal{L}_{TL}$. Figure 4 shows a scheme of the training procedure.

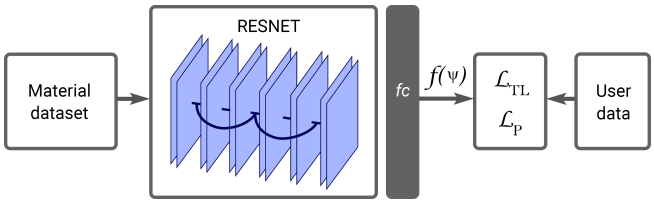

Fig. 4. Scheme of the training process, using both image data from our material dataset, and human data of perceived similarity. We train our model so that, for an input image ψ, it yields a 128-dimensional feature vector $f(\psi)$.

6 EVALUATION

We evaluate our model on the set of images of the material dataset not used during training. We employ the *accuracy* metric, which represents the percentage of triplet answers correctly predicted by our model. It can be computed as *raw*, considering each of the five answers independently as the correct one, or *majority*, considering the majority opinion as correct [Garces et al. 2014; Wills et al. 2009]. Using our MTurk data from Section 4, the results are 73.10% and 77.53% respectively for human observers, indicating a significant agreement across subjects. Our model performs better than human accuracy, with 73.97% and 80.69% respectively. In other words, our model predicts the majority's perception of similarity almost 81% of the time. We include an *oracle* predictor in Table 1, which has access to all the human answers and returns the majority opinion; note that its raw accuracy is not 100 due to human disagreement. Figure 5 shows examples from our 26,000 queries where our model agrees with the majority response, while we discuss failure cases later in this section. More examples of queries and our model's answer are included in the supplementary material.

6.1 Comparison with other metrics

We compare the performance of our model to six different metrics used in the literature for material modeling and image similarity: The three common metrics analyzed by Fores and colleagues [2012], the perceptually-based metrics by Sun et al. [2017] and Pereira et al. [2012], and SSIM [Wang et al. 2004], a well-known image similarity metric. We analyze again accuracy, and we additionally analyze *perplexity*, which is a standard measure of how well a probability model predicts a sample, taking into account the uncertainty in the model. Perplexity Q is given by:

$$Q = 2^{-\frac{1}{|\mathcal{A}|}\sum_\Omega \log_2 p_{ra}} \quad (5)$$

where $\Omega = (r, a) \in \mathcal{A}$, $|\mathcal{A}|$ is the number of collected answers, and p_{ra} is the probability of a being similar to r (Section 5.1). Perplexity gives higher weight where the model yields higher confidence; its value will be 1 for a model that gives perfect predictions, 2 for a model with total uncertainty (random), and higher than 2 for a model that gives wrong predictions. As Table 1 shows, our model captures the human perception of appearance similarity significantly better, as indicated by the higher accuracy and lower perplexity values. Note that perplexity cannot be computed for humans nor the oracle, since they are not probability distributions.

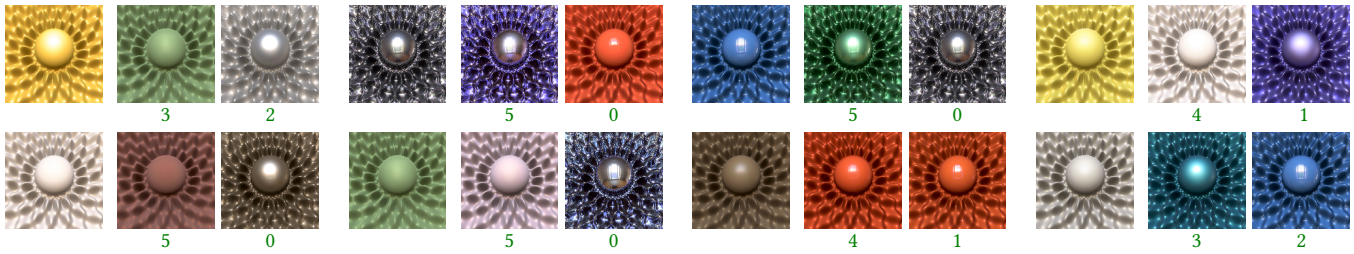

Fig. 5. Examples from our 26,000 queries (reference, plus the two candidates) where our model agrees with the majority response (this is the case almost 81% of the time). The numbers indicate the number of votes each image received from the participants. More examples are included in the supplementary material.

Table 1. Accuracy and perplexity of our model compared to human performance, an oracle (which always returns the majority opinion), and six other metrics from the literature: RMS, RMS-cos, Cube-root [Fores et al. 2012], L2-lab [Sun et al. 2017], L4-lab [Pereira and Rusinkiewicz 2012] and SSIM [Wang et al. 2004]. For accuracy, higher values are better, while for perplexity lower are better.

Evaluation of our model

Metric	Accuracy		Perplexity	
	Raw	Majority	Raw	Majority
Humans	73.10	77.53	-	-
Oracle	83.79	100.0	-	-
RMS	61.63	64.72	3.61	3.13
RMS-cos	61.60	64.67	3.86	3.33
Cube-root	63.71	67.40	1.96	1.86
L2-lab	63.76	67.21	2.16	2.07
L4-lab	60.60	62.93	15.36	11.66
SSIM	62.35	64.74	2.02	1.94
Our model	73.97	80.69	1.74	1.55

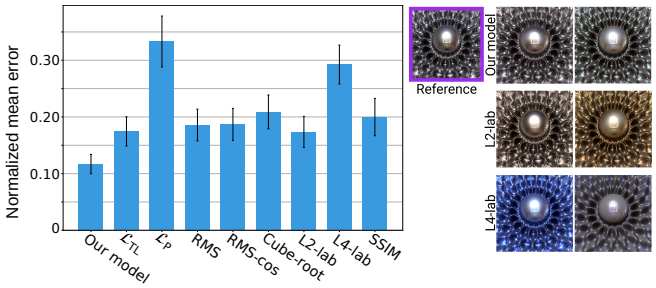

Fig. 6. **Left**: Mean error for different metrics (each normalized by its maximum value) with respect to distances derived from human responses, across all possible pair combinations from the MERL dataset (the $\mathcal{L}_{TL}$ and $\mathcal{L}_P$ columns refer to the ablation studies in Table 2; please refer to the main text). Error bars correspond to a 95% confidence interval. **Right:** Representative example of the two most similar materials to a given reference, according to (from top to bottom): Our model, and the two perceptually-based metrics L2-lab [Sun et al. 2017], and L4-lab [Pereira and Rusinkiewicz 2012]. Our model yields less error, and captures the notion of appearance similarity better.

Additionally, we compute the mean error between distances derived from human responses and our model's predictions, across all possible material pair combinations from the MERL dataset. To obtain the derived distances from the collected human responses, we use t-Distributed Stochastic Triplet Embedding (tSTE) [Van Der Maaten and Weinberger 2012], which builds an n-dimensional embedding that aims to correctly represent participants' answers. We use a value of $\alpha = 5$ (degrees of freedom of the Student-t kernel), which correctly models 87.36% of the participants' answers. We additionally compute the mean error for the six other metrics. As shown in Figure 6, our metric yields the smallest error. Error bars correspond to a 95% confidence interval.

6.2 Ablation study

We evaluate the contribution of each term in our loss function to the overall performance via a series of ablation experiments (see Table 2). We first evaluate performance using only one of the two terms ($\mathcal{L}_{TL}$ and $\mathcal{L}_P$) in isolation. We also analyze the result of incorporating two additional loss terms, which could in principle apply to our problem: A cross-entropy term $\mathcal{L}_{CE}$, and a batch-mining triplet loss term $\mathcal{L}_{BTL}$. The former aims at learning a soft classification task by penalizing samples which do not belong to the same class [Szegedy et al. 2016], while the latter has been proposed in combination with the cross-entropy term to improve the model's generalization capabilities and accuracy [Gao and Nevatia 2018] (more details about these two terms can be found in the appendix). Last, we analyze performance using *only* these two terms ($\mathcal{L}_{CE}$ and $\mathcal{L}_{BTL}$), without incorporating participants' perceptual data. As Table 2 shows, none of these alternatives outperforms our proposed loss function. Although the single-term $\mathcal{L}_P$ loss function yields higher accuracy, it also outputs higher perplexity values; moreover, as Figure 6 shows, the mean error is much higher, meaning that it does not capture the notion of similarity as well as our model.

6.3 Alternative networks

We have tested two alternative architectures, VGG [Simonyan and Zisserman 2014], which stacks convolutions with non-linearities; and DenseNet [Huang et al. 2017], which introduces concatenations between different layers. Both models have been trained using our loss function. As shown in Table 2, both yield inferior results compared to our model. DenseNet has a low number of learned parameters, insufficient to capture the data distribution, hampering convergence. VGG has a larger number of parameters; however, the

Table 2. Accuracy and perplexity for other loss functions, as well as for two alternative architectures (VGG and DenseNet).

Ablation study and alternative networks

Model	Accuracy		Perplexity	
	Raw	Majority	Raw	Majority
$\mathcal{L}_{TL}$	69.32	74.12	1.89	1.73
$\mathcal{L}_{P}$	75.22	82.31	3.16	2.13
$\mathcal{L}_{TL} + \mathcal{L}_{P} + \mathcal{L}_{CE}$	71.82	77.53	1.76	1.66
$\mathcal{L}_{TL} + \mathcal{L}_{P} + \mathcal{L}_{CE} + \mathcal{L}_{BTL}$	71.78	77.76	1.76	1.67
$\mathcal{L}_{CE} + \mathcal{L}_{BTL}$	56.88	58.44	1.96	1.93
VGG	70.70	76.40	2.25	1.89
DenseNet	60.90	63.49	2.66	2.46
Our model	73.97	80.69	1.74	1.55

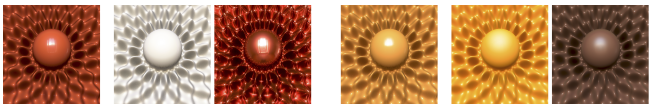

Fig. 7. Two examples where humans' majority disagrees with our metric. For both, humans agreed that the middle stimulus is perceptually closer to the reference on the left, while our metric scores the right stimuli as more similar.

residual mapping learned by the residual blocks in the architecture of our model yields the best overall performance.

6.4 Results by category

We additionally divide the materials into eight categories: *acrylics, fabrics, metals, organics, paints, phenolics, plastics*, and *other*, and analyze raw and majority accuracy in each. We can see in Table 3 how our model is reasonably able to predict human perception also within each category. For instance, although the numbers are relatively consistent across all the categories, humans perform on average slightly worse for phenolics or acrylics, and better for fabrics; our metric mimics such behavior. The only significant difference occurs within the *organics* category, where our metric performs worse than humans. This may be due to the combination of a low number of material samples and a large variety of appearances within such category, which may hamper the learning process.

6.5 Failure cases

Being on par with human accuracy means that our similarity measure disagrees with the MTurk majority 19.31% of the time. Figure 7 shows two examples where humans were consistent in choosing one stimuli as closer to the reference (5 votes out of 5), yet our metric predicts that the second one is more similar. In the leftmost example, the softness of shadows may have been a deciding factor for humans. In the rightmost example, humans may have been overly influenced by color, whilst our metric has factored in the presence of strong highlights. These examples are interesting since they illustrate that neither color nor reflectance are persistently the dominant factors when humans judge appearance similarity between materials.

7 APPLICATIONS

We illustrate here several applications directly enabled by our similarity measure.

7.1 Material suggestions

Assigning materials to a complex scene is a laborious process [Chen et al. 2015; Zsolnai-Fehér et al. 2018]. We can leverage the fact that the distances in our learned feature space correlate with human perception of similarity to provide controllable material suggestions. The artist provides the system with a reference material, and the system delivers perceptually similar (or farther away) materials in the available dataset, thus creating a controlled amount of variety without the burden of manually selecting each material. Figure 1 illustrates this, where the search distance is progressively extended from a chosen reference, and the materials are then assigned randomly to each cube. Suggestions need not be automatically assigned to the models in the scene, but may also serve as a palette for the artist to choose from, facilitating browsing and navigation through material databases. Figure 8 shows two MERL samples used as queries, along with returned suggestions from the *Extended MERL* dataset [Serrano et al. 2016]. The figure shows results at close, intermediate, and far distances from the query. Additional examples can be seen in Figure 9, and in the supplementary material.

7.2 Visualizing material datasets

The feature space computed by our model can be used to visualize material datasets in a meaningful way, using dimensionality reduction techniques. We illustrate this using UMAP (Uniform Manifold Approximation and Projection [McInnes and Healy 2018]), which helps visualization by preserving the global structure of the data. Figure 10 shows two results for the MERL dataset, using images not included in the training set. On the left, we can observe a clear gradient in reflectance, increasing from left to right, with color as a secondary, softer grouping factor. The right image shows a similar visualization using only three categories: *metals*, *fabrics*, and *phenolics*.

7.3 Database clustering

For unlabeled datasets like Extended MERL, our feature space allows to obtain clusters of perceptually similar materials, as shown in Figure 12 using UMAP. The close-ups highlight how materials with similar appearance are correctly grouped together by our model. To further analyze the clustering enabled by our perceptual feature space, we rely on the Hopkins statistic, which estimates randomness in a data set [Banerjee and Dave 2004]. A value of 0.5 indicates a completely random distribution, lower values suggest regularly-spaced data, and higher values (up to a maximum of 1) reveal the presence of clusters. The Hopkins statistic[4] computed over our 128-dimensional feature vectors for the Extended MERL dataset yields a value of 0.9585, suggesting that meaningful clusters exist in our learned feature space (Figure 13 shows three representativve clusters using the Extended MERL database). For comparison purposes, using only *metals* in MERL the Hopkins statistic drops to 0.6935, since

[4]This is an averaged value over 100 iterations, since the computation of the Hopkins statistic involves random sampling of the elements in the dataset.

Table 3. Statistics per category. **From left to right:** Category, number of materials in each category, number of collected answers, humans' accuracy (raw and majority), accuracy of our model, and oracle raw accuracy.

Analysis per material category

Category	Materials	Answers	Humans		**Our model**		Oracle
			Raw	Majority	Raw	Majority	Raw
Acrylics	4	4719	67.27	70.69	67.57	74.18	79.89
Fabrics	14	16019	79.65	83.70	83.03	90.44	87.87
Metals	26	32337	74.20	78.90	75.63	83.10	84.54
Organics	7	8370	69.28	73.08	60.46	62.43	81.28
Paints	14	15101	74.22	78.85	75.22	81.84	84.61
Phenolics	12	13025	66.49	70.53	67.62	74.36	79.72
Plastics	11	12031	70.53	74.70	69.25	74.06	82.05
Other	12	13198	74.80	79.38	78.21	86.11	84.89
Total	100	114800	73.10	77.53	73.97	80.69	83.79

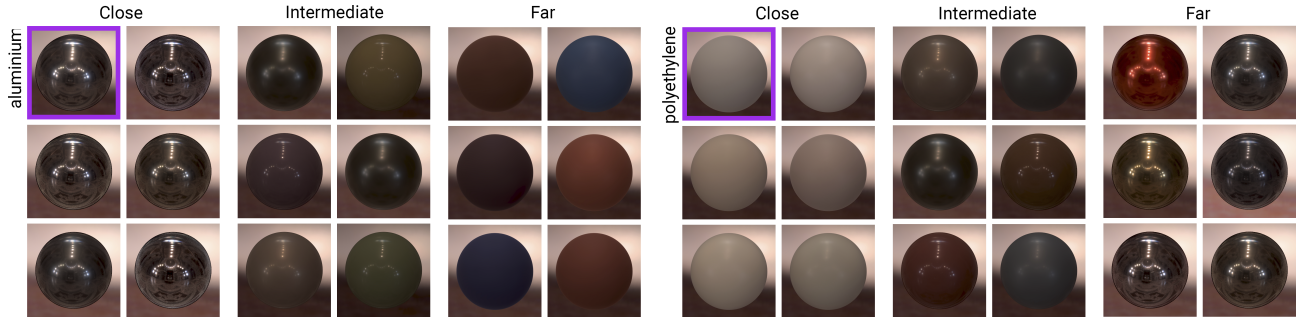

Fig. 8. Two examples of material suggestions using our model. Queries from MERL (violet frame), and returned results for perceptually close, intermediate, and far away materials from the Extended MERL dataset.

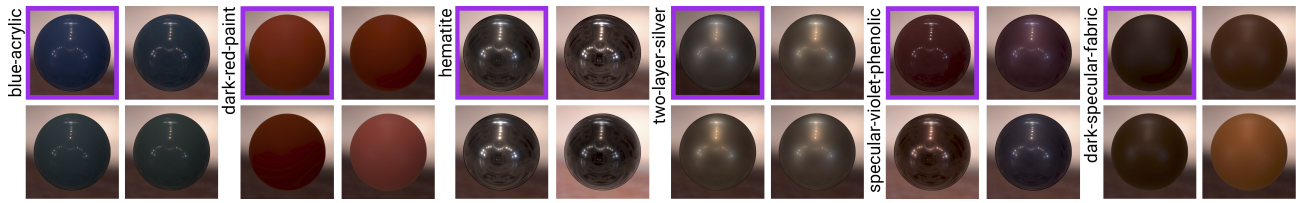

Fig. 9. Additional material suggestion results. Queries (violet frame) and results for the closest materials in the *Extended MERL* dataset.

their visual features are less varied within that category. Figure 11 shows an example of material suggestions leveraging our perceptual clusters in unlabeled datatsets.

7.4 Database summarization

Perceptually meaningful clustering leads in turn to the possibility of database summarization. We can estimate the appropriate number of clusters using the elbow method, taking the number of clusters that explains the 95% of the variance in our feature vectors. In the 400-sample Extended MERL dataset, this results in seven clusters. Taking the closest material to the centroid for each one leads to a seven-sample database summarization that represents the variety of material appearances in the dataset (Figure 14).

7.5 Gamut mapping

In general, our model can be used for tasks that involve minimizing a distance. This is the case for instance of gamut mapping, where the goal is to bring an out-of-gamut material into the available gamut of a different medium, while preserving its visual appearance; this is a common problem with current printing technology, or in the emerging field of computational materials. We illustrate the effectiveness of our technique in the former. Gamut mapping can be formulated as a minimization on image space [Pereira and Rusinkiewicz 2012; Sun et al. 2017]. We can use our feature vector $f(\psi)$ to minimize the perceptual distance between two images as $min_w||f(o) - f(g * w)||_2^2$, where o is the out-of-gamut image, and $g * w$ represents the image in the printer's gamut, defined as a linear

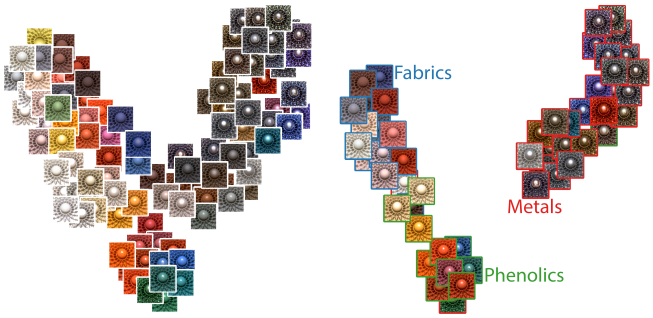

Fig. 10. Visualization of the MERL dataset in a 2D space based on the feature vectors provided by our model, using UMAP [McInnes and Healy 2018]. **Left:** The entire MERL dataset. **Right:** Materials from three different categories (*metals*, *fabrics*, and *phenolics*).

combination of inks g [Matusik et al. 2009]). Figure 15 shows some examples.

8 DISCUSSION

We have presented and validated a model of material appearance similarity that correlates with the human perception of similarity. Our results suggest that a shared perception of material appearance does exist, and we have shown a number of applications using our metric. Nevertheless, material perception poses many challenges; as such there are many exciting topics not fully investigated in this work. Several factors come into play that influence material appearance, i.e., the visual impression of a material, in a highly complex manner; fully identifying them and understanding their complex interactions is an open, fundamental problem. As a consequence of these interactions, the same material (e.g., plastic) may have very diverse visual appearances, whereas two samples of the same material may look very different under different illumination conditions [Fleming et al. 2003; Vangorp et al. 2007]. In aiming for material appearance similarity, we aim for a material similarity metric that can predict human judgements. There is a distinction, common in fields like psychology or vision science, between the distal stimulus—the physical properties of the material—, and the proximal stimulus—the image that is the input to perception—. The key observation here is that human perceptual judgements usually lie between these two, and our training framework and loss function are designed to take both into account. We combine the information about the physical properties of the material contained in the images, by having the same material under different geometries and illuminations, with the human answers on appearance similarity. In other words, a pure image similarity metric would not be able to generalize across shape, lighting or color, while a BRDF-based metric would be unable to predict human similarity judgements.

We do not attempt to identify nor classify materials (Figure 16). Our loss function could, however, incorporate additional terms (such as the cross-entropy and batch-mining triplet loss term discussed in the appendix) to help with classification tasks. We have carried out some tests and found anecdotical evidence of this, but a thorough analysis requires a separate study not covered in this work.

Despite having trained our model on isotropic materials, we have found that it may also yield reasonable results with higher-dimensional inputs. Figure 17 shows three examples from the Flickr Material Database (FMD) [Sharan et al. 2009], which contains captured images of highly heterogeneous materials. We have gathered all the materials from the *fabrics, metals*, and *plastics* categories in the database; taking one reference from each, we show the three closest results returned by our model, using an L2 norm distance in feature space. Images were resized to match the model's input size, with no further preprocessing. Note that the search was not performed within each category but across all three, yet our model successfully finds similar materials for each reference. This is a remarkable, promising result; however, a more comprehensive analysis of in-the-wild, heterogeneous materials is out of the scope of this paper.

We have also tested the performance of our model on grayscale images. In this case, we have repeated the evaluation conducted in Table 1 for our model, using grayscale counterparts of the images. Despite the removal of color information, we obtain results similar to those of our model on color images: A raw accuracy of 72.55 (vs 73.97 on color images), a majority accuracy of 78.64 (vs 80.69), a raw perplexity of 1.82 (vs 1.74), and a majority perplexity of 1.67 (vs 1.55). This further enforces the idea that we learn a measure of appearance similarity, and not image similarity.

To collect similarity data for material appearance, we have followed an adaptive sampling scheme [Tamuz et al. 2011]; following a different sampling strategy may translate into additional discriminative power and further improve our results. Our model could potentially be used as a feature extractor, or as a baseline for transfer-learning [Sharif Razavian et al. 2014; Yosinski et al. 2014] in other material perception tasks. A larger database could translate into an improvement of our model's predictions; upcoming databases of complex measured materials (e.g., Dupuy et al. [2018]) could be used to expand our training data and lead to a richer and more accurate analysis of appearance. Our methodology for data collection and model training could be useful in these cases. Similarly, upcoming network architectures that may outperform our ResNet choice could be adopted within our framework. Finding hand-engineered features could also be an option and may increase interpretability, but it could also introduce bias in the estimation.

In addition to the applications we have shown, we hope that our work can inspire additional research and different applications. For instance, our model could be of use for designing computational fabrication techniques that take into account perceived appearance. It could also be used as a distance metric for fitting measured BRDFs to analytical models, or even to derive new parametric models that better convey the appearance of real world materials. We have made our data available for further experimentation, in order to facilitate the exploration of all these possibilities.

ACKNOWLEDGMENTS

We want to thank the anonymous reviewers for their encouraging and insightful feedback on the manuscript, Ibon Guillen for his invaluable help using Mitsuba; Miguel Galindo, Ibon Guillen, Adrian Jarabo, and Julio Marco for their help setting up the scenes, and

Fig. 11. Material suggestions using our perceptual database clustering. The images show random materials assigned from three different clusters of varying appearance. The robot model (cKalten) was obtained from TurboSquid.

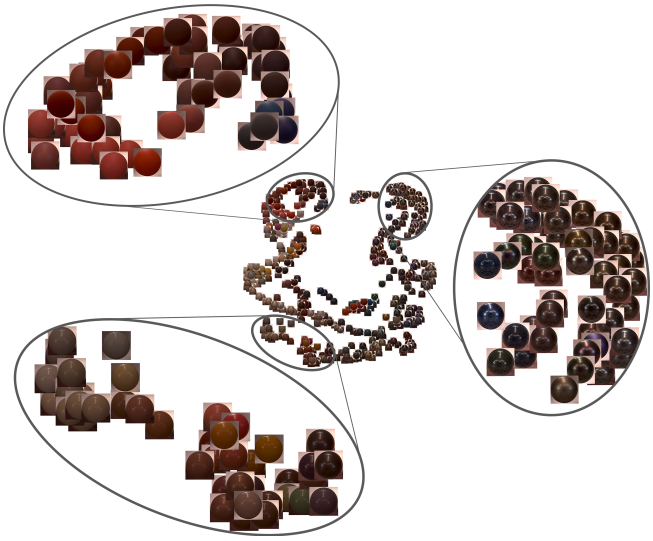

Fig. 12. Visualization of the Extended MERL dataset in a 2D space based on the feature vectors provided by our model, using UMAP. Close-ups illustrate how materials of similar appearance are clearly clustered together. A larger version is included in the supplementary material.

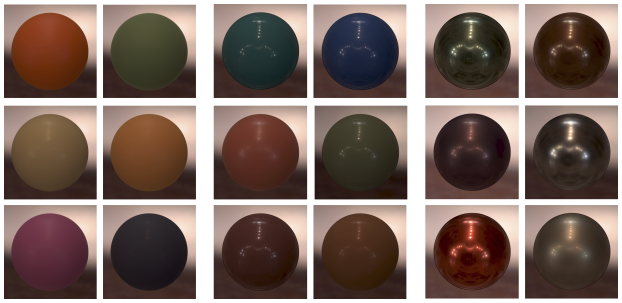

Fig. 13. Representative samples of three clusters on the Extended MERL database. The Hopkins statistic on our feature space confirms that our similarity metric creates perceptually-meaningful clusters of materials.

the members of the Graphics and Imaging Lab for the discussions about the paper. Sandra Malpica was supported by a DGA predoctoral grant (period 2018-2022). Ana Serrano was supported by an

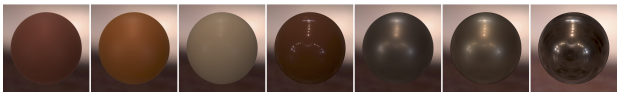

Fig. 14. Example of database summarization for the Extended MERL dataset. These seven samples represent the variety of material appearances in the dataset.

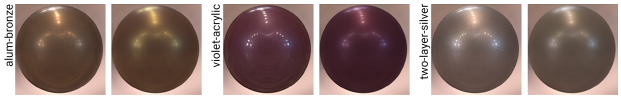

Fig. 15. Our similarity metric can be used for gamut mapping applications, by minimizing the perceptual distance of our feature vectors. Each pair shows the ground truth (left), and our in-gamut result (right).

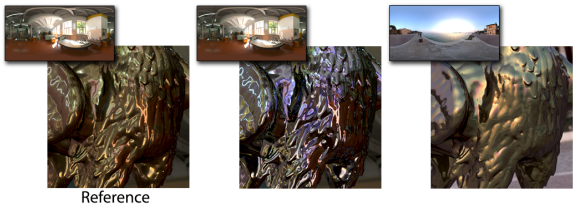

Fig. 16. In the feature space defined by our model, the middle image (*chrome*) is closer in appearance to the reference (*brass*) than the image on the right (*brass*). The insets show the environment maps used. Our model is driven by appearance similarity, and does not attempt to classify materials.

Fig. 17. Results using highly heterogeneous materials from the FMD dataset. We show the three closest results returned by our model, from the reference materials highlighted in violet. Note that the search was performed across all three categories shown, not within each category.

FPI grant from the Spanish Ministry of Economy and Competitiveness, and a Nvidia Graduate Fellowship. Elena Garces is additionally supported by a Juan de la Cierva Fellowship. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project, grant agreement No 682080) and the Spanish Ministry of Economy and Competitiveness (projects TIN2016-78753-P, and TIN2016-79710-P).

REFERENCES

Edward H Adelson. 2001. On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging VI*, Vol. 4299. International Society for Optics and Photonics, 1–13.

Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. 2007. Generalized non-metric multidimensional scaling. In *Conference on artifical intelligence and statistics*. 11–18.

Barton L Anderson. 2011. Visual perception of materials and surfaces. *Current biology* 21, 24 (2011), R978–R983.

Amit Banerjee and Rajesh N Dave. 2004. Validating clusters using the Hopkins statistic. In *2004 IEEE International conference on fuzzy systems (IEEE Cat. No. 04CH37542)*, Vol. 1. IEEE, 149–153.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2013. Opensurfaces: A richly annotated catalog of surface appearance. *ACM Transactions on graphics (TOG)* 32, 4 (2013), 111.

Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. 2015. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 3479–3487.

A.C. Chadwick and R.W. Kentridge. 2015. The perception of gloss: A review. *Vision research* 109 (2015), 221 – 235.

Kang Chen, Kun Xu, Yizhou Yu, Tian-Yi Wang, and Shi-Min Hu. 2015. Magic decorator: automatic material suggestion for indoor digital scenes. *ACM Transactions on graphics (TOG)* 34, 6 (2015), 232.

De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. 2016. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 1335–1344.

D. W. Cunningham, C. Wallraven, R. W. Fleming, and W. Strasser. 2007. Perceptual Reparameterization of Material Properties. In *Proceedings of the third Eurographics conference on computational aesthetics in graphics, visualization and imaging (Computational Aesthetics'07)*. 89–96.

Kristin J. Dana, Bram van Ginneken, Shree K. Nayar, and Jan J. Koenderink. 1999. Reflectance and Texture of Real-world Surfaces. *ACM Transactions on graphics (TOG)* 18, 1 (Jan. 1999), 1–34.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. Ieee, 248–255.

Katja Doerschner, Roland W. Fleming, Ozgur Yilmaz, Paul R. Schrater, Bruce Hartung, and Daniel Kersten. 2011. Visual Motion and the Perception of Surface Material. *Current biology* 21, 23 (2011), 2010 – 2016.

Julie Dorsey, Holly Rushmeier, and François Sillion. 2010. *Digital modeling of material appearance*. Elsevier.

Jonathan Dupuy and Wenzel Jakob. 2018. An Adaptive Parameterization for Efficient Material Acquisition and Rendering. *ACM Transactions on graphics (TOG)*.

Jirí Filip and Radomír Vávra. 2014. Template-based sampling of anisotropic BRDFs. In *Computer graphics forum*, Vol. 33. Wiley Online Library, 91–99.

Roland W. Fleming. 2014. Visual perception of materials and their properties. *Vision research* 94 (2014), 62 – 75.

Roland W Fleming. 2017. Material perception. *Annual review of vision science* 3 (2017), 365–388.

Roland W Fleming, Ron O Dror, and Edward H Adelson. 2003. Real-world illumination and the perception of surface reflectance properties. *Journal of vision* 3, 5 (2003), 3–3.

Roland W. Fleming, Shin'ya Nishida, and Karl R. Gegenfurtner. 2015. Perception of material properties. *Vision research* 115 (2015), 157 – 162.

Adria Fores, James Ferwerda, and Jinwei Gu. 2012. Toward a perceptually based metric for BRDF modeling. In *Color and imaging conference*, Vol. 2012. Society for Imaging Science and Technology, 142–148.

Jiyang Gao and Ram Nevatia. 2018. Revisiting Temporal Modeling for Video-based Person ReID. *arXiv preprint arXiv:1805.02104* (2018).

Elena Garces, Aseem Agarwala, Diego Gutierrez, and Aaron Hertzmann. 2014. A Similarity Measure for Illustration Style. *ACM transactions on graphics (TOG, proc. SIGGRAPH)* 33, 4 (2014).

Stamatios Georgoulis, Vincent Vanweddingen, Marc Proesmans, and Luc Van Gool. 2017. Material Classification under Natural Illumination Using Reflectance Maps. In *IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 244–253.

Ioannis Gkioulekas, Bruce Walter, Edward H Adelson, Kavita Bala, and Todd Zickler. 2015. On the appearance of translucent edges. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 5528–5536.

Ioannis Gkioulekas, Bei Xiao, Shuang Zhao, Edward H Adelson, Todd Zickler, and Kavita Bala. 2013. Understanding the role of phase function in translucent appearance. *ACM Transactions on graphics (TOG)* 32, 5 (2013), 147.

Vlastimil Havran, Jiri Filip, and Karol Myszkowski. 2016. Perceptually Motivated BRDF Comparison using Single Image. *Computer graphics forum* (2016).

Eric Hayman, Barbara Caputo, Mario Fritz, and Jan-Olof Eklundh. 2004. On the significance of real-world conditions for material classification. In *European conference on computer vision (ECCV)*. Springer, 253–266.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.

Yun-Xian Ho, Michael S Landy, and Laurence T Maloney. 2006. How direction of illumination affects visually perceived surface roughness. *Journal of vision* 6, 5 (2006), 8–8.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely Connected Convolutional Networks.. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Vol. 1. 3.

Wenzel Jakob. 2010. Mitsuba renderer. http://www.mitsuba-renderer.org.

Adrian Jarabo, Hongzhi Wu, Julie Dorsey, Holly Rushmeier, and Diego Gutierrez. 2014. Effects of Approximate Filtering on the Appearance of Bidirectional Texture Functions. *IEEE Transactions on visualization and computer graphics* 20, 6 (2014).

Maurice Kendall and Jean D. Gibbons. 1990. *Rank Correlation Methods* (5 ed.). A Charles Griffin Title.

M G Kendall and B Babington-Smith. 1940. On the Method of Paired Comparisons. *Biometrica* 31 (1940), 324–345.

Kihwan Kim, Jinwei Gu, Stephen Tyree, Pavlo Molchanov, Matthias Nießner, and Jan Kautz. 2017. A lightweight approach for on-the-fly reflectance estimation. In *Proceedings of the international conference on computer vision (ICCV)*. 20–28.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Jaroslav Křivánek, James A. Ferwerda, and Kavita Bala. 2010. Effects of global illumination approximations on material appearance. *ACM transactions on graphics (TOG, proc. SIGGRAPH)* 29, 4, Article 112 (July 2010), 112:1–112:10 pages.

Manuel Lagunas, Elena Garces, and Diego Gutierrez. 2018. Learning icons appearance similarity. *Multimedia tools and applications* (2018), 1–19.

Thomas Leung and Jitendra Malik. 2001. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision* 43, 1 (2001), 29–44.

Stephen Lombardi and Ko Nishino. 2012. Reflectance and natural illumination from a single image. In *European conference on computer vision (ECCV)*. Springer, 582–595.

Feng Lu, Xiaowu Chen, Imari Sato, and Yoichi Sato. 2018. SymPS: BRDF symmetry guided photometric stereo for shape and light source estimation. *IEEE transactions on pattern analysis and machine intelligence* 40, 1 (2018), 221–234.

Zhaoliang Lun, Evangelos Kalogerakis, and Alla Sheffer. 2015. Elements of style: learning perceptual shape style similarity. *ACM Transactions on graphics (TOG)* 34, 4 (2015), 84.

Laurence T Maloney and David H Brainard. 2010. Color and material perception: Achievements and challenges. *Journal of vision* 10, 9 (2010), 19–19.

Rafał Mantiuk, Scott Daly, and Louis Kerofsky. 2008. Display adaptive tone mapping. In *ACM Transactions on graphics (TOG)*, Vol. 27. ACM, 68.

Wojciech Matusik, Boris Ajdin, Jinwei Gu, Jason Lawrence, Hendrik P.A. Lensch, Fabio Pellacini, and Szymon Rusinkiewicz. 2009. Printing Spatially-Varying Reflectance. *ACM transactions on graphics (TOG, proc. SIGGRAPH Asia)* 28, 5 (Dec. 2009).

Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. 2003. A Data-Driven Reflectance Model. *ACM Transactions on graphics (TOG)* 22, 3 (July 2003), 759–769.

Brian McFee and Gert Lanckriet. 2011. Learning Multi-modal Similarity. *Journal of machine learning research* 12 (2011), 491–523.

Leland McInnes and John Healy. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

M Mylo, M Giesel, Q Zaidi, M Hullin, and R Klein. 2017. Appearance bending: A perceptual editing paradigm for data-driven material models. *Vision, modeling and visualization. The Eurographics association* (2017).

Addy Ngan, Frédo Durand, and Wojciech Matusik. 2005. Experimental Analysis of BRDF Models. In *Eurographics Symposium on Rendering (2005)*. The Eurographics association.

Addy Ngan, Frédo Durand, and Wojciech Matusik. 2006. Image-driven Navigation of Analytical BRDF Models.. In *Rendering Techniques*. 399–407.

Fabio Pellacini, James A. Ferwerda, and Donald P. Greenberg. 2000. Toward a Psychophysically-based Light Reflection Model for Image Synthesis. In *Proceedings of the 27th annual conference on computer graphics and interactive techniques (SIGGRAPH '00)*. 55–64.

Thiago Pereira and Szymon Rusinkiewicz. 2012. Gamut mapping spatially varying reflectance with an improved BRDF similarity metric. In *Computer graphics forum*, Vol. 31. Wiley Online Library, 1557–1566.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237* (2019).

Michael Rubinstein, Diego Gutierrez, Olga Sorkine, and Ariel Shamir. 2010. A Comparative Study of Image Retargeting. *ACM transactions on graphics (TOG, proc. SIGGRAPH Asia)* 29, 6 (2010), 160:1–160:10.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 815–823.

M. Schultz and T. Joachims. 2003. Learning a Distance Metric from Relative Comparisons. In *Advances in neural information processing systems*.

Gabriel Schwartz and Ko Nishino. 2016. Material Recognition from Local Appearance in Global Context. *arXiv preprint arXiv:1611.09394* (2016).

Gabriel Schwartz and Ko Nishino. 2018. Recognizing Material Properties from Images. *arXiv preprint arXiv:1801.03127* (2018).

Ana Serrano, Diego Gutierrez, Karol Myszkowski, Hans-Peter Seidel, and Belen Masia. 2016. An Intuitive Control Space for Material Appearance. *ACM Transactions on graphics (TOG)* 35, 6, Article 186 (Nov. 2016), 186:1–186:12 pages.

Lavanya Sharan, Ruth Rosenholtz, and Edward Adelson. 2009. Material perception: What can you see in a brief glance? *Journal of vision* 9, 8 (2009), 784–784.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR) workshops*. 806–813.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Cyril Soler, Kartic Subr, and Derek Nowrouzezahrai. 2018. A Versatile Parameterization for Measured Material Manifolds. In *Computer graphics forum*, Vol. 37. Wiley Online Library, 135–144.

Tiancheng Sun, Henrik Wann Jensen, and Ravi Ramamoorthi. 2018. Connecting Measured BRDFs to Analytic BRDFs by Data-driven Diffuse-specular Separation. In *ACM transactions on graphics (TOG, proc. SIGGRAPH Asia)*. Article 273, 273:1–273:15 pages.

Tiancheng Sun, Ana Serrano, Diego Gutierrez, and Belen Masia. 2017. Attribute-preserving Gamut Mapping of Measured BRDFs. *Computer graphics forum* 36, 4 (July 2017).

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2818–2826.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. 2011. Adaptively Learning the Crowd Kernel. In *Proceedings of the 28th international conference on machine learning (ICML'11)*. 673–680.

William Thompson, Roland Fleming, Sarah Creem-Regehr, and Jeanine Kelly Stefanucci. 2011. *Visual Perception from a Computer Graphics Perspective* (1st ed.). A. K. Peters, Ltd., Natick, MA, USA.

Jan Jaap R Van Assen, Pascal Barla, and Roland W Fleming. 2018. Visual features in the perception of liquids. *Current biology* 28, 3 (2018), 452–458.

Laurens Van Der Maaten and Kilian Weinberger. 2012. Stochastic triplet embedding. In *Machine learning for signal processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 1–6.

Peter Vangorp, Pascal Barla, and Roland W Fleming. 2017. The perception of hazy gloss. *Journal of vision* 17, 5 (2017), 19–19.

Peter Vangorp, Jurgen Laurijssen, and Philip Dutré. 2007. The Influence of Shape on the Perception of Material Reflectance. *ACM Transactions on graphics (TOG)* 26, 3, Article 77 (July 2007).

Raquel Vidaurre, Dan Casas, Elena Garces, and Jorge Lopez-Moreno. 2019. BRDF Estimation of Complex Materials with Nested Learning. In *IEEE Winter conference on applications of computer vision (WACV)*.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

Peter Welinder, Steve Branson, Pietro Perona, and Serge J Belongie. 2010. The multidimensional wisdom of crowds. In *Advances in neural information processing systems*. 2424–2432.

Josh Wills, Sameer Agarwal, David Kriegman, and Serge Belongie. 2009. Toward a Perceptual Space for Gloss. *ACM Transactions on graphics (TOG)* 28, 4, Article 103 (Sept. 2009), 103:1–103:15 pages.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks?. In *Advances in neural information processing systems*. 3320–3328.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE Computer Society, 586–595.

Károly Zsolnai-Fehér, Peter Wonka, and Michael Wimmer. 2018. Gaussian Material Synthesis. *ACM Transactions on graphics (TOG)* 37, 4, Article 76 (July 2018), 76:1–76:14 pages.

A ADDITIONAL LOSS TERMS

We describe here the two additional loss terms that we evaluate in our ablation study (refer to Section 6 for details).

A.1 Cross-entropy term $\mathcal{L}_{CE}$

This term accounts for the soft-label cross entropy [Szegedy et al. 2016]. It aims at learning a soft classification task by penalizing samples which do not belong to the same class. In our case, each material represented in the dataset can constitute a class, and the set of classes in the dataset is $\mathcal{K}$. Given an image r included in a training batch $\mathcal{B}$, the probability of r belonging to a certain class $k \in \mathcal{K}$ is given by $p_k(r)$. The cross-entropy loss term is given by:

$$\mathcal{L}_{CE} = \frac{1}{|\mathcal{B}|} \sum_{r \in \mathcal{B}} s(r) \tag{6}$$

$$s(r) = -\sum_{k \in \mathcal{K}} \left[(1-\epsilon) \log p_k(r) l_k(r) + \epsilon \log p_k(r) u(k) \right] \tag{7}$$

where $l(r)$ is the one-hot encoding of the ground truth label, and $l_k(r)$ is the value of the vector for label k (note that our training image data can be labeled, since it comes from the materials dataset presented in Section 3). The value of ϵ is set to 0.1, and we use the uniform distribution $u(k) = \frac{1}{|\mathcal{K}|}$. Both ϵ and $u(k)$ work as regularization parameters so that a wrong prediction does not penalize the cost function aggressively, while preventing overfitting.

A.2 Batch-mining triplet loss term $\mathcal{L}_{BTL}$

In learned models for classification or recognition, a batch-mining triplet loss has been proposed in combination with a soft-label cross entropy term such as the one we use to improve the model's generalization capabilities and accuracy [Gao and Nevatia 2018]. It is modeled as:

$$\mathcal{L}_{BTL} = \frac{1}{|\mathcal{B}|} \sum_{r \in \mathcal{B}} \Big[\operatorname*{argmax}_{x_i^+} \left(||f(r) - f(x_i^+)||_2^2 \right) - \operatorname*{argmin}_{x_i^-} \left(||f(r) - f(x_i^-)||_2^2 \right) + \mu \Big]_+ \tag{8}$$

where x_i^+ designates images of the training batch $\mathcal{B}$ belonging to the same class as r, and x_i^- images belonging to a different class than r. Intuitively, this loss mines and takes into consideration the hardest examples within each batch, improving the learning process.

Received January 2019