Unsupervised Learning Can Predict Object Properties of Non-Rigid Mirror Objects in Ambiguous Conditions

Omer F. Yildiran
École normale supérieure - PSL
Paris, France
omer.yildiran@ens.psl.eu

Katherine R. Storrs
University of Auckland
Auckland, New Zealand
katherine.storrs@gmail.com

Roland W. Fleming
Justus-Liebig-Universität Giessen
Giessen, Germany
roland.w.fleming@psychol.uni-giessen.de

Katja Doerschner
Justus-Liebig-Universität Giessen
Giessen, Germany
katja.doerschner@psychol.uni-giessen.de

Abstract

Although 'glossiness' is an optical property of materials, while 'softness' is a mechanical property, there is an intriguing perceptual connection between the two as both specular reflections and shape deformations produce distinctive motion patterns. Observers are generally excellent at determining properties of moving surfaces. However, under certain circumstances, reflections and deformations can actually be confused, with rigidly transforming mirrors appearing non-rigid and deforming matte-textured objects occasionally appearing somewhat shiny. Here, we investigated whether similar broad successes and specific confusions also arise in an unsupervised recurrent neural network (PixelRNN) trained to predict video sequences. We previously found that such networks reproduce several key gloss perception phenomena, including the 'sticky-reflection' effect [5], wherein reflections that move with the surface (instead of sliding across it) appear like matte texture markings. We generated 8,000 20-frame movies of objects with diverse appearances and motions by varying the shape, glossiness, soft-body properties (including rigid objects), illumination environments, texture maps, and various rotational directions and speeds. After training, the PredRNN-v2 could synthesize accurate video predictions up to 10 frames into the future. T-SNE visualization of the embedding of new stimuli in the network's internal representation revealed a clear clustering and disentanglement of the different object types by the network, as confirmed by logistic linear classification according to their surface glossiness and deformability. Over the course of just the first three frames, the classification of softness increased dramatically from chance to 99% accuracy. These findings demonstrate that unsupervised predictive learning can disentangle the softness and glossiness properties of objects, much like humans, without any explicit training about the distal object properties.

[1]

1. Introduction

Every day, we interact with objects of different kinds, such as soft, rigid, specular, opacity, roughness, wetness, temperature, and so on. We can easily identify

[1]**Note**: It is an unedited draft and not all authors have worked on the article yet

these material properties. Glossiness is one of the fundamental material properties that we perceive in objects. Glossiness can be defined as the amount of light reflected from the surface of materials and how this light appears to our vision. Accurately distinguishing glossy objects from matte objects is a fundamental aspect of visual perception [7, 17]. Previous studies have found that the brain uses a set of simple heuristics to measure proximal properties of gloss in common contexts, rather than estimating the concrete physical properties of surfaces. In other words, the brain measures whatever it can and combines statistical patterns to correlate with glossiness [7]. This property is also particularly interesting as even a tiny modification of an image can change the whole perception of the image. Another important material property in visual perception is the rigidity and softness of materials. Rigidity refers to the material property of resistance to bending, while softness is a material's deformability [19]. These characteristics are strongly linked to the strength of the material, but they differ in that soft, malleable metals can be strong but not rigid, while brittle materials can be rigid but not strong. So distinguishing objects based on their glossiness and rigidity is crucial, as these material properties provide vital information about the objects in our environment. While humans are generally proficient at discerning between these properties, certain visual cues can lead to confusion. In this study, we explore the potential of unsupervised recurrent neural networks to learn and differentiate between glossiness and rigidity, aiming to demonstrate their ability to emulate human-like disentanglement without explicit knowledge of the distal object properties.

The connection between glossiness and softness in terms of perception is intriguing, despite the glossiness being an optical property and the softness being a mechanical one. This connection arises from the distinct motion patterns produced by both specular reflections and shape deformations. Typically, humans excel at discriminating objects based on their rigidity and glossiness properties [11, 22]. However, there are situations where these visual cues can become ambiguous, leading to challenges in accurately judging the material properties of objects.

While some research has been conducted on the topic [1, 8, 18], there is still limited information available on the mechanisms underlying surface reflectance perception. Previous studies have shown that motion measurements can significantly influence perceived material appearance, leading to variations based on image motion characteristics[5]. For instance, specular objects exhibit distorted reflections of the surrounding world, while matte objects display rigidly attached reflections from the environment. Behavioral results from these studies indicated that participants were at chance level when judging the surface gloss of objects, and they tended to perceive objects with motion as shinier. The motion patterns generated by specular reflections were found to depend crucially on surface curvature, indicating that motion plays a crucial role in appearance perception.

To grasp the nature of information essential for assessing the surface reflectance of objects, the researchers employed a systematic approach. They trained a linear classifier on various flow measures derived from sequences in the behavioral experiment, aiming to discern patterns and relationships. The results of this analysis were insightful: the linear discriminant analysis demonstrated that these classifiers excelled in accurately predicting not only the ground truth values but also the participants' performance, offering a robust validation of their findings.

2. Related Work

Our research question was formulated mainly based on the following researches. First of all, a study of Kate Storrs and her colleagues focused on static images and employed unsupervised deep neural networks to explore how visual systems may learn statistical regularities in proximal pictures rather than explicit mappings between proximal cues and known distal causes [20]. They demonstrated that by using an unsupervised generative deep neural network (PixelV) to compress and spatially predict images of surfaces, the network spontaneously clustered inputs by distal factors such as material and illumination, accurately reproducing human observers' misperceptions. The model was established based on assumptions from the surface reflectance literature, where variations in a lower-dimensional group of environmental factors generate variability in proximate sensory inputs, and statistically discernible effects in the proximal input must be generated by diverse distal sources [12].

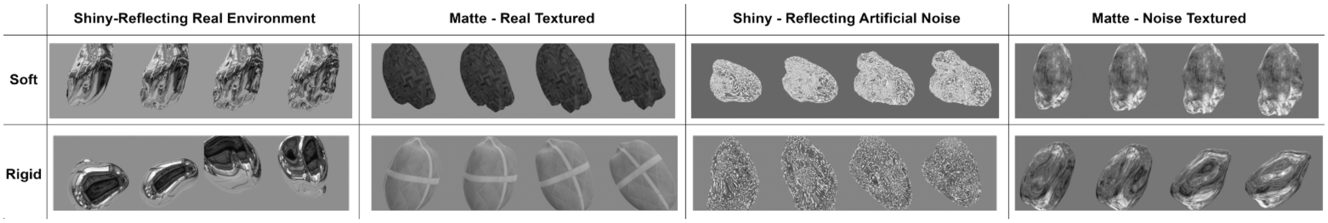
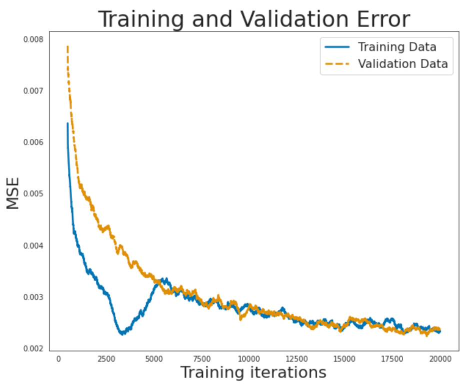
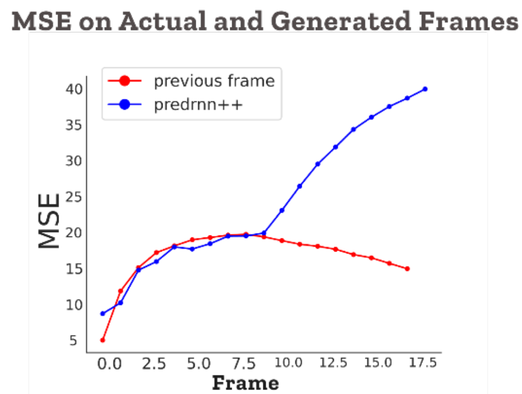


Figure 1. Overview of all eight object types included in the dataset.



(a) Overall mean MSE values.



(b) Example object MSE values between frames.

Figure 2. Training and validation errors between frames across the training iterations(a). Mean Squared Errors between two frame for an example object after 200k training.

For training the neural network, an unsupervised learning technique was chosen over supervised learning due to several reasons. In supervised learning, the network receives explicit labels for outcome data, effectively guiding the machine on where to look in the image. However, supervised learning can be susceptible to weaknesses when dealing with minor image disruptions [9] and may excessively rely on specific textures [10]. Given the complexity and uncertainty surrounding specific gloss perception factors, opting for unsupervised learning was deemed more suitable as it does not rely on predefined labels, allowing the network to autonomously learn and disentangle material properties.

In this research, the PixelVAE neural network was utilized, which facilitated the interpretation of properties and spatially predicted images. This probabilistic graphical model not only enabled property estimation but also generated entirely new pictures with high-order statistical patterns similar to the training data. The material dataset was carefully constructed, consisting of objects with either specular or matte surfaces. Each image was randomly structured with various material properties such as depth, color, and bump intensity. Additionally, the objects were illuminated by six different real-world illumination maps. Ground truth values were utilized for training labels, and the performance of the network was compared with that of other supervised DNN networks, including ResNet. The use of unsupervised learning and the PixelVAE neural network allowed for a comprehensive exploration of material properties in a data-driven manner, overcoming the limitations of supervised learning in handling complex and ambiguous material perception factors. The study's approach provides valuable insights into understanding material properties and their representation in neural networks, contributing to the advancement of computer vision and cognitive science research.

For human rating comparison, various experimental methods were employed, including rating tasks with the original test set and images generated by the DNN model, as well as two alternative forced-choice tasks where participants selected glossier objects. The re-

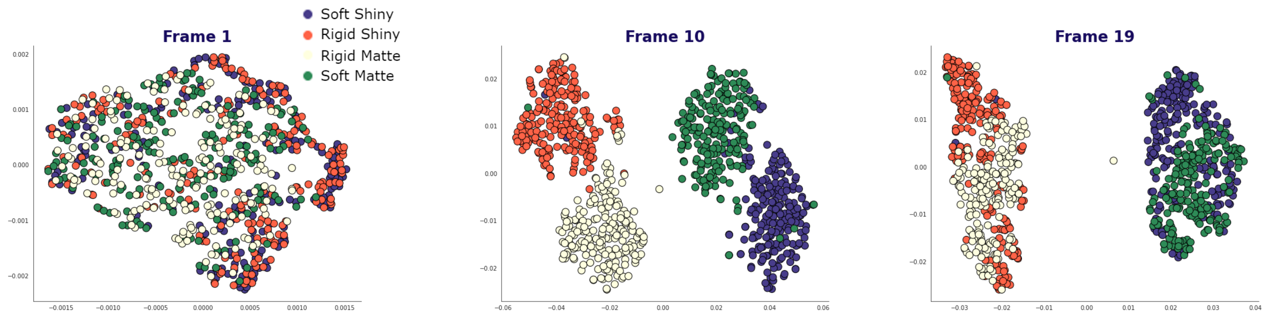


Figure 3. tSNE representations of validation data set over object's softness and glossiness properties on frame 1,10 and 19.

sults showed a high correlation between the model's predicted gloss values and human perception of gloss in images. The study's second aim was to generate plausible new images using pixel sampling with the unsupervised model, showing that the representation of gloss could be better modeled with unsupervised deep neural networks, achieving closer alignment with perceptual rating patterns and reduced error.

In a subsequent study by the same authors [21], they trained an unsupervised neural network, PredRNN-2, on moving objects to investigate if predictive learning could provide useful information for discriminating objects based on their material properties. The model was able to learn up to the 10th frame and predict the next 10 frames based on the previous frames. The network's classifications aligned with human judgments. In this study, we build upon these previous findings and employ an unsupervised recurrent neural network, specifically the PredRNN architecture, to explore the network's ability to disentangle the glossiness and rigidity properties of objects. By leveraging unsupervised predictive learning, we aimed to demonstrate that these neural networks can capture the perceptual connection between glossiness and softness, much like humans, without the need for explicit training on distal object properties. Our investigations pave the way for a deeper understanding of the complex interplay between optical and mechanical properties in visual perception and contribute to the advancement of artificial intelligence models capable of replicating human-like perceptual processes.

3. Method

In line with previous related works, we generated a dataset of $8,000$ **twenty-frame** moving objects using the **Blender 3D Python** pipeline. Each frame of the object consisted of a **256 by 256** pixel image. These objects were carefully varied in terms of shape, illumination environments, texture maps, rotational directions, speeds, and sizes. However, our main focus was on manipulating the surface reflectance and rigidity properties of these objects.

The objects in the dataset were categorized into eight different types based on their surface reflectance and rigidity properties. Specifically, objects were classified as either completely specular or completely matte in terms of their surface reflectance. Regarding rigidity, objects exhibited either orientational motions on three axes or ten different types of soft motions, including squeezing, stirring, and forcing, among others.

3.1. Material mesh

To create the materials, we initially generated primitive spheres with high surface subdivisions using the Catmull–Clark technique [2]. The meshes were then randomly rotated and scaled, enabling us to achieve smoother surfaces and realistic deformations. To deform the shape of the mesh, default noise displacements were randomly applied to the material. The noise scale (ranging from 1 to 2) and noise depth (ranging from 0.8 to 2.0) were also controlled randomly within specified ranges.

3.2. Material surface

For shiny materials, we set properties such as "**metallic**," "**specular**," "**anisotropic**," and "**transmission**" to 100%, while **roughness** was set to 0%. Conversely, for matte materials, we applied the opposite settings.

For matte objects, texture embedding was necessary, and we used a combination of **real-world textures** (10 real-world textures) and **procedural noise textures** (10 noise images). On the other hand, shiny objects did not require texture embedding, as the material already provided a shiny appearance. The surface of these objects reflected only the environmental light, which was captured using either real-world images (10 real-world HDRI images) or procedural noise images.

3.3. Future frame generation accuracy

3.4. Animation

The animations consisted of two types of motion, each lasting **20 frames**. For rigid animation, the object was rotated around three axes, and the rotation angles were randomly generated. For non-rigid animation, the object was given a soft body property, and deformation was achieved using force fields. The force fields were randomly generated from **10 types of built-in force field methods** in Blender, including Drag, Fluid, Guide, Turbulence Charge, Lennardj, Wind, and Force. The Soft Body simulation in Blender simulates the behavior of soft-body objects colliding with other objects. The vertices of the meshes are treated as particles with mass in this environment, and their movement is influenced by the applied forces. The strength, bending, and speed of the force parameters are controlled using random values.

3.5. Noise images

We utilized Perlin noise to create various types of noise images. Ten levels of Perlin noise were generated with different levels of detail(octave), size of each feature(scale), persistance which controlled the influence of octave and lacuranity(frequency multiplier of successive octaves). Each of the 10 noise images was histogram-matched with 10 real-world light probes, resulting in a total of 100 variations of the 10 noise images. Histogram matching enabled us to have a consistent level of pixel luminance and histogram on the overall data set.

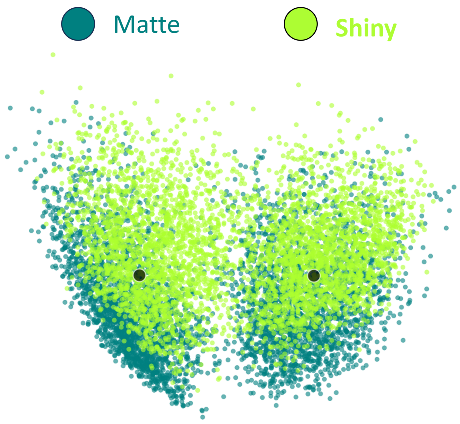
4. Results

The dataset was trained on the PredRNN-v2 recurrent convolutional network, which incorporates the Spatial-Temporal Long Short Term Memory (ST-LSTM) model to preserve spatial and temporal information, enabling the generation of future frames based on past context. The network comprised 4 hidden layers, each with 128 feature maps, and was trained for 200 thousand iterations. The primary objective was to generate one frame into the future based on the 1-10th frame input. As an exploration, the network was also tested without further training after the 10th frame, aiming to generate the future frames until the end of the sequence. To evaluate the network's performance, a separate test set of 800 objects, with 100 samples for each of the 8 types, was created.

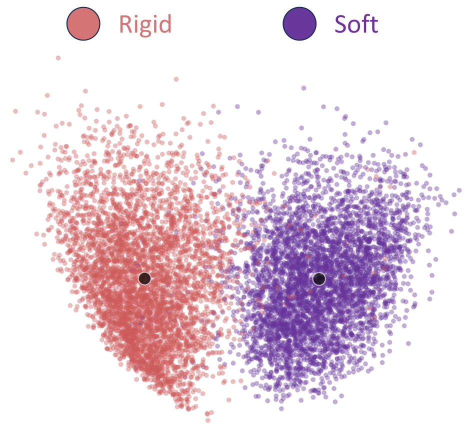
To assess the network's achievement, mean squared errors (MSE) between the frames of objects and their corresponding one-frame future predictions were analyzed. The MSE decreased progressively for both the training and validation data until approximately 175k iterations, reaching a mean MSE of around 0.002. Analyzing the object-specific MSE values at the end of 200k iterations, it was observed that the predicted frame MSE closely aligned with the actual frame MSE up until the 10th frame. The higher MSE difference after the 10th frame was expected, as the goal was to evaluate if the network could still preserve material property information even for frames not included in the training.

4.1. Embedding results

Using various methods, we generated t-distributed stochastic neighborhood embedding (tSNE) visualizations of our test set, which demonstrated the network's successful clustering of objects based on their glossiness and softness aspects. Observing the 4-way t-SNE plots, we noticed that at the first frame, the network could not effectively distinguish an object's softness or glossiness. However, by the 6th frame, the network had already achieved accurate separation, and by the 10th frame, the clustering had significantly improved. After the 10th frame, as the network was not trained further, the glossiness separation accuracy in the em-



(a) Glossiness KMeans plot



(b) Rigidity KMeans plot

Figure 4. KMeans clustering of validation set, seperately colored by rigidity and glossiness properties.

bedding decreased, while the rigidity separation remained accurate. To corroborate the tSNE results, we applied the k-means clustering algorithm to the overall objects, achieving 99% accuracy in differentiating objects based on rigidity, but only 55% accuracy in glossiness clustering.

4.2. Classification Results

To obtain more robust results, we employed linear classification using the Logistic Linear Classifier on internal representations of both the training and test sets. This partial supervision of the model allowed for a more comprehensive evaluation. The classification accuracy at each frame was highly accurate for glossiness, even at the first frame, while it was at chance level for rigidity. However, after just 4 frames, rigidity classification surpassed glossiness classification, achieving 99.9% accuracy. The almost perfect classification accuracy for each object property rendered this analysis less informative for evaluating with human participants. Thus, we focused on the network's classification decision values for glossiness and rigidity properties.

The decision values for both factors were normally distributed across the objects. For glossiness classification, only two objects were misclassified as Shiny-Soft instead of Matte-Soft. However, the decision value analysis did not provide a clear answer regarding

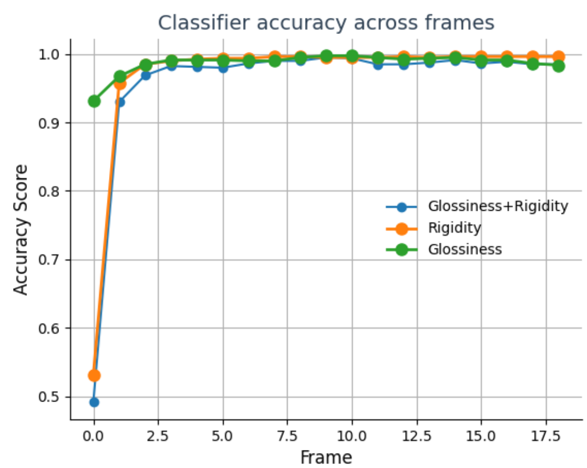
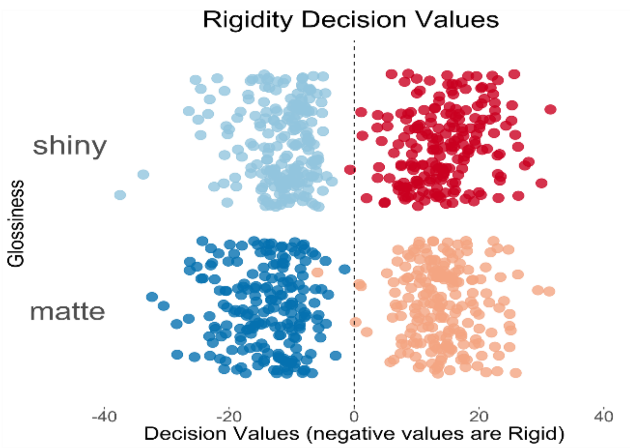
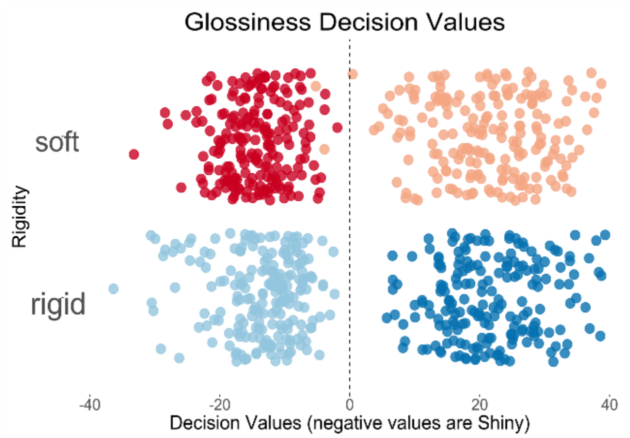


Figure 5. Classification accuracy across the frames for 4-way classification(blue), rigidity classification (orange) and glossiness classification (green).

the cues used by the network to classify these objects. In Figure 6, we present the decision values obtained for each object in the test set. Figure 6a (on the left) displays the decision values derived from the classifier trained on the ground truth rigidity type of each object. The x-axis of this plot represents the decision values, while the y-axis illustrates the division based on two categories: shiny and matte. Within each category, the objects' positions are intentionally jittered vertically to enhance visual separation of the data points, providing



(a) Rigidity decision values



(b) Glossiness decision values

Figure 6. Classification decision values for each object based on rigidity classifier (a) and glossiness classifier(b)

a clearer understanding of the distribution.

Vice-versa, Figure 6b (on the right) illustrates the decision values generated by the classifier trained on the physical glossiness category of each object. The y-axis in this plot is divided into two distinct categories: soft and rigid, representing the objects' rigidity levels. The decision values along the x-axis offer insights into the classifier's predictions, indicating how well it distinguishes between soft and rigid objects based on their physical glossiness characteristics.

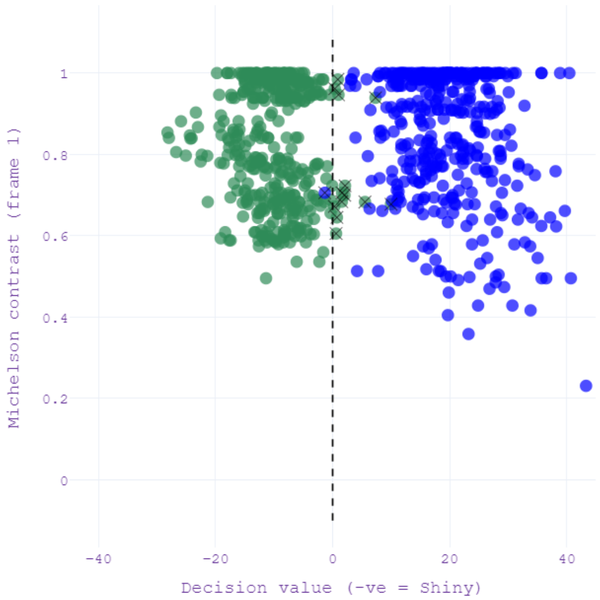
By looking at Figure 6a and Figure 6b, it is possible to analyze the classifier's performance in classifying objects based on physical glossiness and rigidity types. The misclassified objects in both of these figures are very close to a decision value of zero which explains that when a classifier makes a mistake in classification, our classifier is not sure about its prediction decision.

To delve deeper into the decision values, we plotted contrast values as a function of decision values. These contrast values were also shaping a uniform distribution across the decision values given on a specific training of the classifier. Meaning that for all of the decision value points contrast values of objects were distributed similarly. Contrary to our prediction, we have stated that for lower contrast values decision value of objects should more likely be lower compared to higher contrasted objects. So we could not conclude with any causal relationship between image contrast and the glossiness classification of an object.

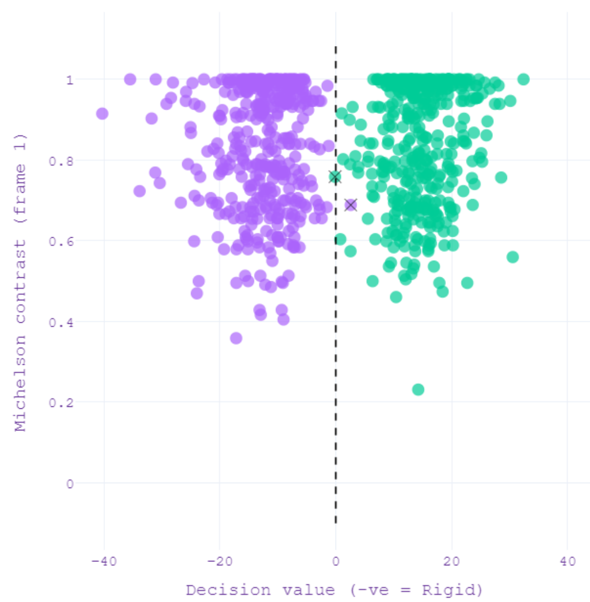
We also created a comprehensive control test set including another type of object. In this class of objects, materials were matte but different from other matte objects these objects were embedded with a texture of real-world HDRI environment images. For humans basically, it would be almost impossible to distinguish a static object that's reflecting a real-world environment than a matte object textured with a real-world object. We first used the same old training set but with a new validation set and checked the classification capabilities for a single frame, the result was not surprising, glossiness accuracy was at chance level. However when we included all other frames of these objects, the logistic regression classification results showed that our network still could achieve 98.2% accuracy, meaning that motion cues could significantly improve the ability to distinguish between glossy and matte objects.

5. Discussion

The results obtained from training the PredRNN-v2 recurrent convolutional network are indicative of its robust learning capabilities. The gradual decrease in MSE for both the training and validation data underscores the network's proficiency in predicting future frames accurately up to the 10th frame. The observed higher MSE difference after the 10th frame was in line with our expectations, confirming the network's abil-



(a) Glossiness contrast values as a function of decision values of a classifier trained on glossiness factor



(b) Rigidity contrast values as a function of decision values of a classifier trained on rigidity factor

Figure 7. Objects' Michelson contrast values with relation to decision values, which the logistic linear classifier is trained for physical glossiness values(a) and also a classifier trained for physical rigidity values(b)

ity to preserve material property information even for frames beyond the training data.

The classification results demonstrate that unsupervised neural networks can effectively learn the glossiness and rigidity properties of moving objects without the need for additional dataset labeling. The classifier based on the internal representations of these objects achieved an impressive accuracy of 99.97% in classifying objects' glossiness and rigidity properties. This high classification accuracy suggests that our network possesses a human-level ability to distinguish between matte and shiny objects, even in ambiguous situations, such as when classifying a matte object with a real-world textured appearance. On the classification aspect, our results provided empirical data to the literature which shows that motion cues can give more information about object properties which aligns with existing studies done on human judgements[4–6]. This also revealed that unsupervised learning can be used to achieve human-level ability in the classification of objects. After all, there are further research needs to be done to compare visual perception with human judgments and model classification decision accuracy values. Furthermore, the t-SNE visualizations and KMeans clustering results provide compelling evidence of the network's ability to cluster objects based on glossiness and softness aspects. The progressive improvement in clustering accuracy over frames suggests that the network learns to differentiate between materials effectively as it receives more temporal context. A notable discovery in our research was the robustness of rigidity classification. The tSNE plots and rigidity classification results indicated that the network could preserve information about rigidity properties even without access to further frames beyond the 10th frame, while the glossiness clustering was losing its stability beyond the trained data. This finding suggests that rigidity might be a more stable property of objects compared to glossiness.

The network's ability to classify objects' glossiness was not significantly affected by contrast analysis. This result was contrary to previous assumptions and the literature which found that high contrast can enhance the perceived objects' glossiness[3, 13, 14]. Our analysis of the network's decision values exhibited invariance to the contrast levels when making glossiness-related decisions.

While our study provides valuable insights, it has limitations. The dataset's scope, though substantial, may not cover all real-world scenarios. Future research could explore expanding the dataset to include a broader range of materials and environmental conditions. Although we already created 2 other versions of the dataset of 8.000 objects using 20 more real-world environment light maps, further examinations are warranted. Also, we created all of the objects using a gray scale for simplicity but using 3 or 4 channel color information might also have impact on the perception of glossiness perception[7]. In summary, our results contribute empirical evidence to the literature, demonstrating that motion cues provide valuable information about object properties for unsupervised neural networks. Moreover, these networks can attain human-level proficiency in judging object properties. However, further research is required to compare human judgments with network classification decision accuracies, providing deeper insights into the network's capabilities and potential areas for improvement.

These implications of our findings may not only contribute to the advancement of computer vision but also hold potential for applications in various fields, including robotics, augmented reality, and autonomous systems where understanding and accurately perceiving object properties are essential for real-world applications. As the field of unsupervised learning continues to advance, our research provides valuable insights into the potential of neural networks to mimic human-like perception and understanding of material properties in dynamic scenes.

References

[1] Edward H Adelson. Lightness perception and lightness illusions. *Perception*, 24(10):1143–1156, 2000. 2

[2] Edwin Catmull and Jim Clark. Recursively generated b-spline surfaces on arbitrary topological meshes. In *Seminal graphics: pioneering efforts that shaped the field*, pages 183–188. 1998. 4

[3] Bin Chen, Chao Wang, Michal Piovarči, Hans-Peter Seidel, Piotr Didyk, Karol Myszkowski, and Ana Serrano. The effect of geometry and illumination on appearance perception of different material categories. *The Visual Computer*, 37: 2975–2987, 2021. 8

[4] Katja Doerschner, Huseyin Boyaci, and Laurence T Maloney. Estimating the glossiness transfer function induced by illumination change and testing its transitivity. *Journal of Vision*, 10 (4):8–8, 2010. 8

[5] Katja Doerschner, Roland W Fleming, Onur Yilmaz, Paul R Schrater, Benjamin Hartung, and Daniel Kersten. Visual motion and the perception of surface material. *Current biology*, 21(23): 2010–2016, 2011. 1, 2

[6] Katja Doerschner, Ozgur Yilmaz, Gizem Kucukoglu, and Roland W Fleming. Effects of surface reflectance and 3d shape on perceived rotation axis. *Journal of vision*, 13(11):8–8, 2013. 8

[7] Roland W Fleming. Human perception: Visual heuristics in the perception of glossiness. *Current Biology*, 22(20):R865–R866, 2012. 2, 9

[8] Roland W Fleming, Ron O Dror, and Edward H Adelson. Real-world illumination and the perception of surface reflectance properties. *Journal of vision*, 3(5):3–3, 2003. 2

[9] Robert Geirhos, Charles R Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in neural information processing systems*, 2018. 3

[10] Robert Geirhos, Katherina Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. In *Advances in Neural Information Processing Systems*, pages 13890–13902, 2020. 3

[11] Benjamin Hartung and Daniel Kersten. Distinguishing shiny from matte. *Journal of Vision*, 2 (7):551–551, 2002. 2

[12] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 2

[13] Hiroaki Kiyokawa, Tomonori Tashiro, Yasuki Yamauchi, and Takehiro Nagai. Manipulation of glossiness perception by contrast enhancement of high spatial frequency components. *Journal of Vision*, 20(11):364–364, 2020. 8

[14] Hiroaki Kiyokawa, Tomonori Tashiro, Yasuki Yamauchi, and Takehiro Nagai. Spatial fre-

quency effective for increasing perceived glossiness by contrast enhancement. *Frontiers in Psychology*, 12:625135, 2021. 8

[15] Anna Metzger and Matteo Toscani. Unsupervised learning of haptic material properties. *Elife*, 11:e64876, 2022.

[16] Isamu Motoyoshi, Shin'ya Y Nishida, Ladan Sharan, and Edward H Adelson. Image statistics and the perception of surface qualities. *Nature*, 447(7141):206–209, 2007.

[17] Takuma Murakoshi, Tomohiro Masuda, Ken Utsumi, Kazuo Tsubota, and Yuji Wada. Glossiness and perishable food quality: visual freshness judgment of fish eyes based on luminance distribution. *PLoS One*, 8(3):e58994, 2013. 2

[18] Shin'ya Y Nishida and Michiko Shinya. Use of image-based information in judgments of surface-reflectance properties. *JOSA A*, 15(12): 2951–296, 1998. 2

[19] Lucie Preißler, Bianca Jovanovic, Jörn Munzert, Filipp Schmidt, Roland W Fleming, and Gudrun Schwarzer. Effects of visual and visual-haptic perception of material rigidity on reaching and grasping in the course of development. *Acta Psychologica*, 221:103457, 2021. 2

[20] Katherine Storrs and Roland Fleming. Learning to see material from motion by predicting videos. *Journal of Vision*, 21(9):1993–1993, 2021. 2

[21] Katherine R Storrs, Barton L Anderson, and Roland W Fleming. Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10):1402–1417, 2021. 4

[22] Gordon Wendt, Franziska Faul, Vebjörn Ekroll, and Rainer Mausfeld. Disparity, motion, and color information improve gloss constancy performance. *Journal of vision*, 10(9):7–7, 2010. 2