Reviewer #1 (Comments for the Author (Required)):

Authors assessed computational luminance constancy with AMA algorithm, with naturalistic images generated by computer graphics tools. It was interesting approach. However, some of critical information to understand the approach seemed to be missing or less comprehensive. It would be great if authors could address those issues.

Thank you for the careful reading and helpful comments. Please see below for how we have clarified in response. Reviewer comments are in black, responses in red.

Luminance constancy was mentioned as "constitutive component of ... general color constancy". However, the definition was not formally given. Authors could have provided the background, any of their specific definition, logic, concept and any assumptions in more details in Introduction.

Thanks for the suggestion. In the original submission, we provided a definition of the computational problem of luminance constancy in the last paragraph of the introduction (Line 58). We have now added a parallel definition of the more general color constancy problem at the start of the second paragraph of the introduction (Line 29):

> “The computational problem of color constancy may be framed as how to obtain stable descriptions of the spectral surface reflectance functions of the objects in a scene.”

(p.3, para 4) "We define the computational problem of luminance constancy as that of estimating the light reflectance value (LRV) of a target object's surface reflectance function. Estimating the LRV from a surface reflectance function proceeds in two steps. First, one computes the luminance of the light that would be reflected from the surface under a reference illuminant. Second, one normalizes the result by the luminance of the reference illuminant itself."

A problem of conventional color constancy is that we do not know the surface reflectances. However, authors' approach seemed that they assume that the reflectances are already available to viewers, independent of illuminants. Those assumptions, if any, and the links to the computations, estimation of the task-optimal receptive fields with cone-excitations and their normalizations, could have been more clearly explained.

As described in response to a comment by Reviewer 2 below, we have corrected "light reflectance value (LRV)" to "luminous reflectance factor (LRF)" throughout.

You are correct that human and computational observers do not generally have information about surface reflectance when viewing novel scenes. Our work only makes use of ground truth information about surface reflectance in the construction and design of our computational observer. When we test the observer’s ability to estimate LRF on images of novel scenes, the computational observer does not have any more information about the surface reflectance than would in principle be available to a human observer viewing the same scenes. When evaluating performance, the observer never has direct access to any quantity other than the cone-responses.

The passage quoted by the reviewer describes how LRF is defined, rather than how our computational observer estimates LRF. To prevent confusion, we have changed the third sentence in the paragraph from: “Estimating the LRF from a surface reflectance function proceeds in two steps.” to (Line 60):

> “Obtaining the LRF from a known surface reflectance function requires two steps.”

Without those explanations, it is difficult to follow the computations and their results.
e.g. "...datasets to determine how well target object LRV can be estimated from cone excitations and from normalized cone contrasts. Studying both representations allows us to understand how early contrast coding and normalization affect luminance constancy. We applied accuracy maximization analysis (AMA) to learn the optimal receptive fields for estimating LRV, and evaluated the performance obtained when the responses of these receptive fields are optimally decoded."

We hope our response to the previous comment addresses this issue.

Authors introduced the concept of the light reflectance values (LRV) as a "specific problem of luminance constancy, as constitutive component of the more general color constancy problem". However, those "problems" were not well identified in Introduction.

Please see above; we have now expanded our set of definitions.

The relationship of the LRV in a physical world could be clarified.

We have now clarified our definition of LRF in the last paragraph of the introduction, as described above. We think the prose in the paper itself is now sufficiently clear, and have not added additional complexity to Figure 1.

Figure 1 could have improved and to be used to explain the LRV and "object-extrinsic factors".
In Introduction, the property of LRV seemed to be part of physical properties. However, the LRV was one of the parameter in the computation, as if it is one of the internal properties (within visual system).

Indeed, LRF plays two distinct roles in the paper. As stated in the introduction, the LRF is a physical property of object surface reflectance. It is also the quantity that we seek to estimate.

One of the authors has publications about the illumination geometry and its importance. Mutual reflections, shadow, specularit and multiple illuminations are also important in color constancy. Authors could have commented how these properties were considered in the present model.

Thanks for the suggestion. We now specifically speak to these aspects of the scene and rendering process (Line 92):

> "All surfaces in the scene model were matte and did not have specularities. There were multiple light sources in the scene, and the rendering process simulated shadows as well as mutual reflection of light between nearby surfaces."

We did, however, examine the impact of secondary reflections. See next response.

(p.14, last para) "the secondary reflections have minimal effect on LRV estimation: the estimates without secondary reflections were similar to those with reflections."
Does this mean that the computational luminance constancy with AMA cannot address the mutual reflection or the mutual reflection has no effect on the constancy?

We intended the latter. We have reworded the paragraph to increase clarity. It now reads (Line 291):

> "Our rendering software allows us to compare the effect of background surface reflectances on target object LRF with and without simulation of secondary reflections of light from one object onto another. These secondary reflections were included in the dataset from which we report our primary results. When we turn off this feature of the rendering, we find (data not shown) that LRF estimation performance is essentially unchanged. Estimates with and without secondary reflections are very similar. This result suggests that the primary source of the estimation error in Condition 3 is caused by direct effects of image-to-image variation in the reflectance of the background objects on the AMA responses."

Performance of luminance constancy was discussed briefly in Discussion with RMSE. The definition of the relative RMSE and how it could evaluate the luminance constancy was not given in the main text.

We have added the definition of "relative RMSE" at the end of the methods section. It now reads (Line 219):

> "We quantified the performance of AMA and the baseline methods at estimating LRF with relative root mean squared error (relative RMSE). Relative RMSE is the square root of the mean of the squared difference between the estimated and true LRF divided by the true LRF. The mean is taken over all stimuli in the test set."

As for technical matters, there seemed to be any restrictions in using the AMA. Such disadvantages of the computations adopted in the present study could be identified, as well as the advantages. The experimental setting up and parameters could have been explained in more details. An essential point could be what the task or optimization criterion in the AMA was.

Sorry for the confusion. The task that AMA was trained and tested on was to estimate LRF. We have clarified that AMA cost functions we used evaluated the quality of the LRF estimates, in the second paragraph of the methods subsection titled "Learning optimal receptive fields" (Line 186):

> “In our implementation of AMA, we used both the Kullback-Leibler divergence cost function (corresponding to the maximum a posteriori estimator) and the mean squared error cost function (corresponding to the posterior mean estimator) to evaluate the accuracy of the AMA estimates of LRF. Training with both cost functions yielded similar estimation performance; the results reported here are for when the Kullback-Leibler divergence cost function was used in training.”

The spatial resolution of the image data and area seemed to be very small: e.g. "the target by cropping the rendered images to 1 x 1 degrees of visual angle around the target object (51 x 51 pixels)". Authors could provide justifications whether these sizes are large enough to evaluate the effect of LRV, interaction of the geometry of the object surfaces and multiple illuminations.

The choice of a 1°x1° analysis area was informed by data on the size of receptive fields in early visual cortex (Gattass et al, 1981; Gattass et al, 1988). Our thinking is that the computations required for constancy likely have a cortical component. At the same time, keeping the area relatively small allows us to evaluate the information that might reasonably be expected to be integrated within primary visual cortex. Figure 9b shows the image area that was used, and this area contains multiple object surfaces.

That said, we think this point is worth discussing. We have expanded in methods (Line 147) and returned to the point in the second paragraph of the future directions section of the discussion (Line 379):

> Line 147: “We focused our analysis on image regions local to the target by cropping the rendered images to 1°x1° of visual angle around the target object (51 x 51 pixels; Figure 9b). The local analysis is motivated by the fact that neural receptive fields early in the visual pathways (e.g., retina, primary visual cortex) pool information locally. In primary visual cortex, foveal receptive fields have a maximum spatial extent of approximately 1 degree of visual angle (Gattass, Gross, & Sandell , 1981; Gattass, Sousa, & Gross , 1988). We sought to understand how well responses from AMA-learned receptive fields at a similar scale could be used to achieve luminance constancy.”
>
> Line 379: “We studied the information available for LRF estimation using a 1°x1° image patch. As noted above this choice of size was motivated in part to use a spatial scale roughly commensurate with the scale of information integration in early visual cortex. Our general methods could be extended to study larger regions, and doing so would quantify the value of spatially remote information for luminance constancy.”

It is unclear how 51 pixels correspond to 1 degree in visual angle. Thus, the parameters for the simulations were not fully explained; thus, corresponding physical size, distance, direction of light sources, intensity,

The rendering software operates in arbitrary distance units – any overall scaling of the scene and distance to the eye does not change the rendered image. Similarly, we compute optical blur using a PSF that is a function of visual angle (or equivalently, mm on the retina) and thus our simulation of blur depends only on the angular extent of the stimuli. Since the computed image is thus invariant with respect to overall scaling, pixel quantization together with the size of the patch in visual angle provide a complete specification of the relevant stimulus parameters.

Definition of "naturalistic" should be given. For example, a sphere or Xylophone in the air is not seen in everyday life.

We are reminded of a well-known US Supreme Court decision, which we rephrase here as "we can't define naturalistic precisely but we know it when we see it." Indeed, we don't have a sharp definition of naturalistic, but we think the intended meaning – "like natural images but with some remaining artificialness" – is sufficiently clear from the context in which we use the term.

The simulated images were based on the indoor structure, but authors applied outdoor illuminants. The simulations of the illuminants were based on the Granada natural illuminants.

Natural lighting often comes in through windows. We view the mixing and matching referred to by the reviewer as one of the reasons are scenes are naturalistic rather than natural.

Thus, despite using the "natural" dataset, those were decomposed and fitted with linear combination of Gaussians. This may sounds as if "natural" data was transformed to "unnatural".

The spectra were fitted with linear models that provided a highly accurate approximation of the measured spectra. It was the distribution of the coefficients that were modeled as Gaussian distributions, not the spectra or the spectral components. Figure 6 is our best attempt to convey how the distribution of illuminants we used approximates measured natural spectra.

Some terms and acronym should be defined and explained at the first appearance.
e.g. LRV and the definition of the "relative RMSE" were given in the Fig 14 caption.

The LRF acronym is defined in the abstract (Line 6) and the introduction (Line 58). We have added the definition of “relative RMSE” at the end of Methods, as noted in the response above.

[Methods]
(p.5, second from the last para) "The LRV values were equally spaced between 0.2 and 0.6. For each LRV value, we generated a different relative target object surface reflectance for each scene."

The range of LRV was [0.2 0.6]. What was the meaning of this range? What is the meaning of the LRV 0 and 1?

The range [0.2 0.6] was chosen because the LRF of most (>90%) of the generated surface spectra fell within this range. We now mention this in the text (Line 106):

> “More than 90% of the surface reflectance spectra (generated as described below) fell within this range.”

The meaning of LRF 0 and 1 is now provided in the text (Line 65):

> “An LRF of 0 means that none of the light from the reference illuminant is reflected from the surface. An LRF of 1 means that the luminance of the surface reflectance under the reference illuminant is the same as the luminance of a perfect reflector under that same illuminant.”

(p.5, last para) "The Library base scene contains 2 area lights. We inserted one additional spherical light source into the scene. The position and size of the inserted object, the inserted light source, and the viewpoint on the scene were held fixed across all scenes. "

What was the rationale to use the two area lights?

In many natural scenes, light comes from more than one location.

How these multiple lights were manipulated in the Conditions 2 and 3?

We have added the following to explain the manipulation of the light spectra for Condition 2 and 3 (Line 109).

> “In Conditions 2 and 3, the overall intensities of the three light source illumination spectra were equal, while their relative shape varied. The overall intensity varied from scene to scene.”

(p.4, para 1) "The package builds..."
It would be useful for readers if authors could inform the system requirements and any practical restriction you may be aware of.

The github repositories for the code (links provided in the paper) provide these details. We think it is most useful to keep this sort of information with the code, rather than laying it out in the paper itself.

[Baseline methods]
Why was the 3 x 3 pixels region used? Was it center of the 51 x 51 pixels?

We wanted a region that was within the target object, and indeed the region was at the center of the 51 x 51 pixel image. The choice of 3 x 3 was somewhat arbitrary.

L:M:S ratio was 6:3:1. Does this mean that it was possible that no S-cone was included depending on the area?

There were 277 S cones in the mosaic. We are now explicit about the number of cones of each type (Line 155):

> “The cone mosaic contained L:M:S cones in approximately the ratio 0.6:0.3:0.1 (1523 L cones, 801 M cones, 277 S cones) …”

We used a demosaiced version of the cone-responses, so all 3 types of cones were present in the analysis. This is mentioned in Line 160. The demosaicing process does not add additional information.

Figure 10
(b) What is the meaning of the negative values on x-axis?

The receptive fields contain both positive and negative values. Responses can be positive or negative. That is, we are not modeling the spike rates of real neurons, only the information carried by linear receptive fields. The information signaled by a linear receptive field could be equivalently signaled by one 'On' receptive field and one polarity-reversed 'Off' receptive field, both of which rectify their responses.

(c) What was the spatial dimension of the RF?

The RFs are specified over the same 1 x 1 degree patch as the stimuli. This is now stated explicitly in the caption for Figure 10.

Were the computations of the RF independent across L, M, and S?

Each receptive field is a spatio-chromatic linear filter that was determined jointly by AMA. That is, the L, M and S components of each RF were determined jointly and are not independent in that sense.

Did they have the same spatial size?

Yes. We have now made this clear where we give the size of the RFs in the caption for Figure 10.

[Typos]
p.2, last sentence: "these factors (?, ?; Brainard...)"

We have corrected this.

Figure 10 (a): no "filled region" in the panel.

For this case, the filled region is too small to be visible. We have now expanded the caption to say: "The filled regions representing standard deviations are too small to be visible"

Reviewer #2 (Comments for the Author (Required)):

The authors investigate how differences in object relative reflectance spectrum (i.e. color but not albedo), illumination spectrum and reflectance spectrum of the background affects performance of an optimal decoder in predicting a measure of the surface reflectance (light reflectance value - LRV, which is the reflected luminance under a reference illuminant normalized by the luminance of the illuminant itself - thus, being conceptually similar to albedo).

Specifically, as a first step they generate a large set of rendered naturalistic images systematic varying LRV. In addition to LRV changes, reflectance spectra of the target images, of the illumination and of the background surfaces was varied, according to three conditions. 1) The relative reflectance spectrum of the surfaces was varied while keeping the illumination spectrum and the background reflectance constant, 2) the reflectance and the illumination spectra were varied while keeping the background constant, and 3) all the three factors varied. From condition 1 to 3 estimating LRV from the pixel images is a harder problem because of the additional confounding variations.

As a second step, they used a model of the early visual system to mimic the optical blurring of the eye and the spatial sampling of the three classes of cones. The simulated cone excitations in response to the pixels at the corresponding sampled positions were transformed into images by demosaicing via linear interpolation. Then, the three L M S excitation images were normalized to equate the response magnitude across classes. Cone contrast images were computed from the normalized cone excitations, and both excitation and contrast images were separately used in the analyses.

As a third step, the authors used accuracy maximization analysis (AMA) to determine a set of linear filters (weighting functions applied to the L M S images) chosen to best classify LRV. The AMA searched over the space of linear filters to find the ones that minimize a given cost function. These linear filters are the optimal receptive fields for decoding LRV.
As a final step, they tested how well the responses of these optimal receptive fields can be decoded to estimate LRV. As a baseline, they used predictions form a linear regression fit of LRV as a function of the cone excitation and contrast from a central region of the target images. The receptive fields and the regression coefficients were estimated on the 90% of the images and tested on the remaining 10%.

In condition 1 performance of both AMA and linear regression was close to perfect, based on cone excitation. This is not surprising because only LRV is changing, yielding to a monotonic (linear) increase of cone excitation. In fact, receptive fields are characterized by random weights in the background regions and high weights corresponding to the target regions. This is true for the L and M images, but receptive field applied to S excitation images present a random distribution of small weights, indicating poor contribution of S cones. This is interesting because cone excitation were normalized before the analyses.

In condition 2, based on cone excitation AMA performance was rather poor, reflecting the additional complexity introduced by changing illumination spectrum (thus affecting luminance of the target object). Regression performance was as bad as guessing the mean LRV of the training set. When based on cone contrast, AMA performance was again nearly perfect, and regression improved, presumably because background luminance information is implicit in the images because of the normalization procedure. The shape of the receptive fields was not reported.

In condition 3, performance was only evaluated based on cone contrast images. Both AMA and linear regression performed worse than in condition 2 (AMA performed better than linear regression), reflecting the increased complexity, however they provided information about LRV. The shape of the optimal receptive fields revealed systematic contribution of cone excitations from background object locations, with positive and negative relatively high weights from different background regions, presumably because of the correspondence to background objects.

I think the approach presented in the manuscript is interesting because 1) simulations through physically accurate renderings allows to generate databases large enough for statistical learning, and 2) AMA allows to assess what information is relevant, by investing how manipulation affects performance, but also by looking at the structure of the receptive fields. Also, the structure of the receptive fields, will depend on the geometry of the scene (since it was held fixed), thus it can tell where relevant information is (e.g. in the background objects).

Thank you for this concise summary and positive evaluation of our work. Responses to comments are in red below.

In only recommend to improve clarity. In particular, I think confusion is made between definitions of constancy. I think the use of "luminance constancy" is at least misleading if not a contradiction in terms. The authors refer to a normalized luminance measure (LRV), which is close to albedo. I think this needs to be made clear from the title. Also, I do not understand why LRV is chosen rather than albedo, since also albedo can be computed based on luminance and to my knowledge it is a more common magnitude in perception research and computer graphics. I think this choice needs to be commented. If as I think, LRV conceptually corresponds to albedo, I recommend changing "luminance constancy" with "lightness constancy", since lightness is commonly referred to as the perceptual correlate of surface albedo.

This is a reasonable point, and the question of terminology is one we grappled with as we wrote the initial draft. We agree that we did not sufficiently explain the reasoning for our terminological choices.

In the literature, "lightness constancy" generally refers to studies where the stimuli are restricted by be achromatic, and this is not the case we study. At the same time, we are not tackling the full-color case. So, we introduce the term "luminance constancy" to describe specifically the problem we studied. This is now clarified in a footnote where we define luminance constancy (Line 58):

Similarly, although "albedo" is related to what we called "light reflectance value (LRV)", it is not the same. In particular, albedo does not incorporate the human luminosity function. We think we should use the term that most accurately describes what we are doing. That said, as we re-checked the literature we realized that we should have called this "luminous reflectance factor (LRF)". We have made this change throughout.

We have added clarifying footnotes where we introduce luminance constancy and LRF (Lines 58 and 60):

> “We define the computational problem of luminance constancy[1] as that of estimating the luminous reflectance factor (LRF) of a target object's surface reflectance function. The LRF is a measure of the overall amount of light reflected by a surface relative to the reference illuminant itself.[2]”
>
> [1]In the literature, the term lightness constancy is generally used to denote color constancy for the special case when stimuli are restricted to be achromatic. This condition does not apply to our work – we consider full spectral variation in the stimuli. We chose the term luminance constancy to denote the generalization

> from achromatic stimuli. At the same time we acknowledge that we are not studying the full problem of color constancy. Rather, we are studying the estimation of a luminance-based summary of surface spectral reflectance.
>
> [2]LRF is related to albedo, but the concept of albedo does not incorporate the human luminosity function.

Minor comments:

Reference to previous work might be extended, especially concerning research in perception. In fact, although in my understanding, the presented work is about lightness constancy, there is no definition of lightness and it is not clear what are the factors involved in lightness constancy. For a definition of lightness and brightness I recommend referring to "Lightness Perception and Lightness Illusions"- Adelson, 2000. For the factors contributing to lightness constancy I suggest "Seeing black and white" - Gilchrist, 2006.

We modified the end of the first paragraph to add a definition of lightness constancy and to include some key references (Line 24):

> "The ability of a visual system to compute a representation of object color that is stable against variation in object-extrinsic factors is called color constancy. A well-studied special case of color constancy is when the stimuli are restricted to be achromatic. This special case is called lightness constancy (Gilchrist, 2006). Although human lightness and color constancy are not perfect, they are often very good (Foster, 2011; Brainard & Radonjic, 2014; Adelson, 2000; Kingdom, 2011)."

Also, there is a certain body of work on which scenes aspects potential cues for lightness (e.g "Cues to an Equivalent Lighting Model" Boyaci, Doerschner & Maloney, 2006; "Illumination estimation in three-dimensional scenes with and without specular cues" - Snyder, Doerschner & Maloney). Specular reflections are one of those cues. However, there are human and simulation studies reporting that specular highlight are discounted in lightness judgments and that specular reflections potentially impair lightness discrimination (e.g. "Lightness constancy in the presence of specular highlights" - Todd, Normal & Mingolla, 2004; "Lightness perception for matte and glossy complex shapes", Toscani, Valsecchi & Gegenfurtner, 2017; "The effect of gloss on perceived lightness" - Beck, 1964).

We agree that providing a bit more in the way of pointers into the relevant literature will be helpful, although reviewing this literature is beyond the scope of the current paper. We have now edited the following paragraph in the discussion, and added citations along the lines suggested above. Changes to this paragraph also address the three points made by the reviewer that follow this one (Line 361).

> "In the work presented here, we studied computational luminance constancy in virtual scenes with naturalistic spectral variation in light sources and in surface reflectance functions, with only matte surfaces in the scenes. It is natural to start with spectral variation, because this variation is at the heart of what makes luminance constancy a rich computational problem. In natural scenes, however,

> there are other sources of variation that add additional richness. These include variation in non-spectral properties of lighting and objects in the scene. Examples include lighting geometry, object texture, object material (e.g. specularity), and object shape. The methods we developed here may be generalized to study the effects of variation in these factors; one could incorporate these factors into the generation of the scenes and again learn estimators from the corresponding labeled images. A challenge for this approach will be to thoughtfully control the increase in problem complexity, both to keep compute time feasible and to ensure that it is possible to extract meaningful insight from the results. Extending the work to include variation of material may provide insights not only about luminance constancy but also for computations that relate to material perception (see Fleming, 2017) Extending the work to include variation in object shape and lighting geometry may clarify the role of object boundaries versus object interiors for providing information that supports perception of object color and lightness (see Land & McCann, 1971; Rudd, 2016). We also note that there is a literature on how increasing stimulus complexity along the various lines listed above affects human color and lightness perception (e.g. Beck, 1964; Yang & Maloney, 2001; Yang & Shevell, 2002; Todd et al., 2004; Snyder et al., 2005; Boyaci et al., 2006; Xiao & Brainard, 2008; Kingdom, 2011; Xiao et al., 2012; Anderson, 2015; Toscani et al., 2017), and the computational problem of color and lightness constancy (e.g. Lee, 1986; D'Zmura & Lennie, 1986; Funt & Drew, 1988; Tominaga & Wandell, 1989; Barron & Malik, 2012; Barron, 2015; Finlayson, 2018)."

I think that the approach presented in the manuscript might help investigating the role of specular highlights for an ideal observers. In fact, with a fixed geometry of the scene and the illumination (as it was in the reported simulations) the distribution of the weights in the receptive fields is informative about the role of the elements in the scene. Given the interest that specular reflection received by color and lightness constancy investigations, I would add this in the "Future Directions" section.

Good point. We have adopted this suggestion in the revised paragraph above.

Also, I suggest stating that the rendered scenes were matte in the "Images of Virtual vs. Real Scenes" section, as a limitation of the simulation given that specular reflections might interact with lightness constancy, as discussed in the literature.

This restriction is now noted explicitly as described above, albeit in the future directions rather than virtual vs. real scenes section.

Classical ("Lightness and retinex theory", Land & McCann, 1971) but also recent theories of lightness constancy ("A cortical edge-integration model of object-based lightness computation that explains effects of spatial context and individual differences" - Rudd, 2014) propose that visual system spatially integrates the luminance steps corresponding to reflectance edges (as given object boundaries). By looking at the shape of the receptive field in condition 3, it seems that rather large positive weights are flanked by negative weights corresponding to borders between the objects in the background, suggesting edge related computations. I suppose one of the potentiality of the approach is to reveal such local computations, thus if the authors find my speculation sensible, I would add it in the discussion, showing how the proposed approach has the power to reveal lightness constancy computations as proposed in the literature.

This is an interesting connection, which we now make explicit, again in the same revised future directions paragraph quoted above. (We cited more recent 2016 paper by Rudd, though.)

The idea of generating large datasets of rendered surfaces in order to investigate classification of an ideal observers (ROC and linear classification) on their material properties is not new ("Optimal sampling of visual information for lightness judgments" - Toscani, Valsecchi & Gegenfurtner 2013; "Lightness perception for matte and glossy complex shapes" - Toscani, Valsecchi & Gegenfurtner 2017; "Statistical correlates of perceived gloss in natural images" -Wiebel, Toscani & Gegenfurtner, 2015).

However, to my knowledge this is the first time that reflectance spectra are taken into account, as opposed to grayscale images, as the toolbox presented in the paper allows. I would stress the novelty respect to previous work.

Thanks for pointing us towards the work above, which indeed we should have cited. We have now remedied this and a few other related omissions through modifications to the following paragraphs in the discussion (Lines 327 and 334).

> “Large datasets of natural or posed scenes with ground truth information about illuminant and object surface reflectance are difficult to obtain, as independent measurement of illuminant and surface properties at each image location is painstaking. That noted, there are databases for evaluation of color constancy algorithms that provide information about illumination and/or surface reflectance (e.g., Barnard et al., 2002; Ciurea & Funt, 2002; Cheng et al. 2014; Nascimento et al., 2016; see http://colorconstancy.com). Often the illumination is estimated through placement of a reflectance standard at a few image locations to allow estimation of the illumination impinging at those locations. These illumination estimates are then interpolated/extrapolated across the image. However, the quality of this approximation cannot typically be evaluated.
>
> Here we used labeled images rendered from descriptions of virtual scenes. A similar approach has been used previously to study the perception of lightness and specularity (Toscani et al., 2013; Weibul et al., 2015; Toscani et al, 2017; Proket etl al., 2017). Our work adds to this approach by introducing color variation. There are many advantages to using rendered images (Butler, Wulff, Stanley, & Black, 2012). One advantage is that they allow us to work with large number of labeled images where object reflectance is precisely known at each pixel. A second advantage is that we can control the variation in distinct scene factors that might affect the difficulty of the estimation problem. This flexibility allows the impact of scene factors to be studied individually or in combination. Here, we exploited this flexibility to quantify how variation in the relative reflectance spectrum of the target object, the spectrum of the illumination, and the reflectance spectrum of the background objects limit LRF estimation. We

> also exploited our use of rendered images to explore how the presence or absence of secondary reflections from background objects affected estimation of target object LRF. This type of question cannot be addressed using real images. The basic approach we use here can be extended to include parametric control over the amount of variation of different factors. For example, we could systematically vary the variances of the distribution over the weights that control the relative spectrum of the illumination."

The distribution of surface albedos in natural environments is approximated by a specific beta distribution ("The distribution of reflectances within the visual environment", Attewell & Baddeley, 2006) and the discernible colors present in nature only cover a specific portion of theoretical solid of visible colors ("The number of discernible colors in natural scenes" Linhares, Pinto & Nascimento, 2008). For the simulation presented in the manuscript, reflectance spectra are sampled from a statistical model approximating a largely variable set of colors, as the Munsell chips is supposed to represent the space of visible colors rather than resembling the occurrence of colors in the word. I suppose this gives an upper limit to the limitation in performance due to the increasing complexity with conditions, and results might change considering the natural distribution of reflectance spectra.

There are databases providing a large collection of reflectance spectra or reflected spectra from isolated surfaces under a known illuminant, although they do not span color spaces as well as the munsell system. In fact, they focus on leaves fruits and vegetables ("Fruits, foliage and the evolution of primate colour vision" - Regan, Julliot, Simmen, Vienot, Charles-Dominique & Mollon, 2011; "Hyperspectral database of fruits and vegetables" - Ennis, Schiller, Toscani & Gegenfurtner, 2018).

We have added discussion of these and related papers into the section of the paper that relates to our approximation to naturally occurring surface reflectances (Line 351).

> "To increase the representativeness of our rendered images, we used datasets of natural surface reflectance spectra and natural daylight illumination spectra. Although we believe these datasets provide reasonable approximations of the statistical variation in reflectance and illumination spectra, they can be extended and improved. For example, there are additional datasets of measured surface reflectances that could be incorporated into future analyses. Some of these datasets focused on the reflectance of objects (e.g. fruit) that are thought to be important for the evolution of primate color vision (e.g., Sumner & Mollon, 200; Regan et al, 2001; Barnard et al., 2002; Ennis et al., 2016). Another issue, not addressed by these datasets, is relative frequency of different surface reflectances in natural viewing. Attewell & Baddeley (2007) performed a systematic survey, and reported the distribution of an LRF-like quantity in natural scenes. Generalizing these measurements to better characterize the distribution of full reflectance functions remains an interesting goal."

I found only one typo at the end of page 2: "(?, ?; Brainard and Freeman, 1997)", probably due to the reference manager.

Fixed.

I would find interesting to have the shape of the receptive fields reported also for the analysis about the scenes in condition 2.

We have added these to the appendix.