Qeios

Peer Review

Review of: "MLI-NeRF: Multi-Light Intrinsic-Aware Neural Radiance Fields"

Ruggero Pintus[1]

1. Visual and Data Intensive Computing Group, CRS4 Centro di Ricerca, Sviluppo e Studi Superiori in Sardegna, Pula, Italy

General assessment

The paper introduces a novel method for extracting intrinsic image components from scenes captured under various camera poses and lighting conditions. The authors propose MLI-NeRF (Multiple Light Information Intrinsic-aware Neural Radiance Fields), an extension of NeRF that incorporates light position and leverages multi-light and multi-view information to enhance intrinsic decomposition performance. The key innovation is using NeRF to generate pseudo-intrinsic reflectance and shading, which guide the training of intrinsic image decomposition without requiring ground truth data. The method has been evaluated on both synthetic and real-world datasets.

The first general observation is that the proposed method employs a NeRF-like network to extract the scene's geometry, represented in this case as a Signed Distance Function (SDF). This geometry is then used to compute the normal and shadow maps for each camera pose, which, combined with known lighting conditions, are utilized to estimate the Lambertian albedo of the surface. These components (SDF, reflectance, and shading) are integrated with the original images to train a network that refines the standard NeRF by incorporating constraints derived from the reflectance and shading intrinsic images. This refinement enhances the performance of novel view synthesis (NVS) and relighting under novel light positions while also enabling shading edits within novel views.

Computing normal maps, shading maps, and albedo from multi-view acquisitions with known light positions is a well-studied and effectively solved problem using various approaches outside of NeRF-like methods. Therefore, the strength of this approach lies in Stage 2, where intrinsic images are leveraged to produce a NeRF that excels in novel view synthesis (NVS) and relighting. In fact, Stage 1 could be replaced with any geometry reconstruction method capable of providing a dense surface

 doi.org/10.32388/B3VSX3

representation from multi-view data. Steps A, B, and C could then be derived from that geometry—whether it is an SDF, point cloud, or mesh—along with the known light positions. The resulting output would still be pseudo-reflectance and shading map pairs that can serve as input to the intrinsic-aware NeRF network. The key contribution of the method, in my view, is demonstrating that a NeRF trained with reflectance and shading constraints (and light position as input) can deliver improved results when rendering novel views under novel lighting conditions.

From this perspective, I would suggest reorganizing the paper, particularly the results section, to emphasize this aspect. Currently, the results focus primarily on intrinsic decomposition, but I would recommend placing greater emphasis on the NVS and relighting performance by comparing outputs for unseen camera poses and lighting conditions with out-of-training images. Comparing this method, which explicitly utilizes known light positions, with approaches that handle unknown lighting (e.g., TensorIR, InvRender, IntrinsicNeRF) is not entirely fair or sound. Knowing the light positions and geometry makes it significantly easier to extract shading and albedo, whereas these other methods aim to solve the problem without light information. Furthermore, the model used here is relatively simple and based on a straightforward Lambertian model. Similarly, I would avoid comparing the intrinsic image decomposition results with methods that do not leverage both multi-view and light information (e.g., PIE-Net and Ordinal), as these methods operate under different assumptions and constraints.

I would recommend shifting the focus of the validation from intrinsic decomposition to the NVS and relighting performance. For example, while NRHints also incorporates light position as a network input (similar to this paper), it does not use a preliminary estimation of reflectance and shading to guide the training. This distinction should be emphasized more prominently throughout the paper, from the related work section to the results and discussions. The validation should prioritize this aspect, with comparisons—both numerical and visual—focusing primarily on the method's ability to improve NVS and relighting under novel light positions and camera views.

In summary, the paper shows promise but requires a complete reorganization in terms of presentation and discussion. Certain sections need to be revised or removed, while others should be rewritten to improve clarity and focus. Additionally, a significant number of new results should be included to strengthen the paper’s contribution.

With that in mind, I’ve outlined a series of detailed comments below.

Related Work

The related work section should be restructured, rewritten, and expanded to provide a more in-depth discussion of the novelty of the proposed approach in relation to the current literature. Many of the statements used to argue that the existing literature lacks exploration of the topic or the application of advanced techniques to more complex scenarios are too general and, in their generality, inaccurate. Below, I have listed all the sentences from the related work section that compare the proposed method with existing literature.

- "...Our approach leverages 3D information and physical constraints (e.g., variations in illumination) to achieve superior results..." ⇒ This statement is too general. How exactly does the approach utilize 3D information and physical constraints, and how does it differ from existing methods in the literature?
- "...However, the potential of using information from multiple lights for 3D scene understanding remains unexplored..." ⇒ This statement is too general and inaccurate, as it fails to acknowledge existing methods that already explore this aspect. It also does not clarify the specific contribution of the current method.
- "...However, inverse rendering methods are primarily focused on individual objects and are difficult to extend to large, complex scenes, such as those with backgrounds..." ⇒ This statement is too general and inaccurate. It does not explain how the proposed method addresses this limitation or contributes to advancing the field.

In general, the authors should be more specific about the problem they aim to address and clearly explain how the proposed method solves that particular problem. They should also provide a more detailed discussion of how their work advances the state-of-the-art by comparing the class of cited works with the novel contributions presented in the paper. I elaborate on this general comment in the following section.

I would suggest adding a section dedicated to pure NeRF, as the paper has three main components: NeRF, Intrinsic Image Decomposition, and Relighting. While two of these components are discussed in the related work, NeRF is not covered. Perhaps inverse rendering could be mentioned only in the introduction, as the presented method does not rely on any inverse rendering approach.

I would recommend avoiding general statements like "...However, the potential of using information from multiple lights for 3D scene understanding remains unexplored....". Such a statement is

inaccurate in its generality, as many methods already exploit multiple light sources for 3D scene understanding. For example, outdoor photometric stereo uses multiple sunlight positions to extract surface normal maps. Please be more specific, especially when discussing the novelty of the approach proposed in the paper.

Similarly, this statement is also inaccurate: "...However, inverse rendering methods are primarily based on individual objects and are challenging to extend to large, complex scenes, such as those with backgrounds...". Many methods exist that address inverse rendering in complex scenes, such as indoor environments with multiple objects, heterogeneous backgrounds, and intricate global illumination.

Overall, the discussion of the novelty of the proposed approach in relation to the existing literature is quite limited. For each section, or perhaps as a final standalone section (e.g., "Our Contribution"), I would suggest providing a more thorough discussion of the novelty of the proposed method, highlighting its advantages and drawbacks compared to the related work. This would help the reader better contextualize the presented work within the state-of-the-art.

For example, the related work section does not clearly indicate whether adding light position to NeRF is a novel contribution. The paper states, "...we extend NeRF to incorporate light position..." (Introduction), or "...we introduce light position as input to extend NeRF..." (Fig. 1). However, reference [7] has already used light position as input for their relighting neural radiance field. In the related work section, reference [7] is cited briefly with the statement, "...Zeng et al. [7] enhanced NeRF relighting with visibility and specular hints...," which fails to communicate that the introduction of light position into NeRF is not a new concept.

Given that the system in Fig. 1 involves a complex procedure, the authors should provide a clear explanation of which elements in the pipeline are novel, as well as the rationale behind their inclusion. This should not only involve describing what is new or not, but also the motivation for choosing these particular solutions and how they improve upon existing approaches.

As a minor comment, I would suggest adding a general reference to inverse or differentiable rendering, as reference [8] is relevant to both intrinsic image decomposition and its close connection to inverse rendering. For example, I would include the following:

- Kato, Hiroharu, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. "Differentiable rendering: A survey." arXiv preprint arXiv:2006.12057 (2020).

Method

A design choice worth considering is why the authors decided to train the MLP-RGB, MLP-Reflectance, and MLP-Shading in parallel, with the constraints applied only during volume rendering, rather than during the production of the RGB image. Specifically, MLP-Reflectance and MLP-Shading could have been placed before MLP-RGB, so they would serve as additional inputs to the MLP-RGB network. In that case, the constraints on the RGB image would have been applied and learned earlier in the process, rather than relying solely on volume rendering. This is important because significant errors in the RGB image cannot be fully corrected by volume rendering alone. By imposing the constraints earlier, it could have been more effective. This approach is somewhat similar to (or follow the same rationale of) Stage 1, where the MLP-RGB network also depends on the output of the SDF network, which learns geometry and enforces that constraint on the MLP-RGB output.

The authors mention that they use multiple light sources to better address indirect illumination and shadows, stating, "...In previous inverse rendering methods [10], accounting for indirect illumination and shadow has been a critical challenge. Our strategy is to leverage multiple lighting conditions to make the corresponding information more accessible...." While they use SDF to extract the normal map, shading (including self-shadows), and light visibility (including cast shadows), indirect illumination is a more complex phenomenon. It is unclear how the paper addresses this issue using multiple lights. Additionally, in the results (e.g., Fig. 1), it is evident that the final outcome is influenced by indirect light from the background, which is not handled by the proposed pipeline. The authors should rephrase some sentences to clarify these concepts, as statements like "...to make the corresponding information more accessible..." might lead the reader to believe that indirect illumination is being considered. However, there is no evidence that global illumination effects, such as interreflections, have been managed or addressed in the paper.

The sentence "...By calculating $R = I \oslash S$, we can compile the reflectance information for every pixel from multiple light conditions while diminishing the influence of the entangled indirect light...." appears to be incorrect. By computing the ratio between the original image and the shading, when the shading is non-zero due to cast shadows, the authors essentially remove the component associated with Lambertian shading under direct lighting. However, this operator does not address or correct the effects of indirect illumination, which will remain entangled. Additionally, the direct lighting will still be partially entangled, as some points may not be perfectly Lambertian. In this case, the operator does

reduce the effect of direct lighting by attempting to estimate the proper surface albedo, but it does not fully disentangle both the direct and the indirect illumination.

Using K-means to compute the albedo is one possible approach, though not necessarily the best. More robust solutions can be applied through trimming methods. For example, after ordering the reflectance values, one could discard the darkest and brightest pixels, and assuming that 50% of the remaining measurements are not outliers in the Lambertian model, the average or median of those remaining 50% could be used. This is just one example, but there are many other approaches. The authors should explain their choice of K-means, providing a rationale for this design decision, perhaps by discussing it in the related work section or, ideally, briefly in the corresponding paragraph.

Please provide a more detailed and numerical explanation of how the weight maps WR and WS are computed. First, include the specific formula used to assign the weights, as only qualitative descriptions have been provided so far, such as “...higher pseudo shading values...”, “...further from visibility edges...”, “...closer to visibility edges...”, or “...with lower pseudo shading values...”. Second, it is unclear whether WR and WS are the same, as the criteria are often described without specifically referring to one weight map or the other. These details are crucial for ensuring reproducibility.

Results

Although the results shown in Fig. 1 are quite promising, including the reflectance and shading editing and simulation, the authors should provide a more detailed discussion of some of the results in Fig. 1. For example, in Fig. 1c, the effect of interreflections (yellow signal) on the zebra’s legs is clearly visible, leading to a deformation of the reflectance. According to the model presented in Section 3.1, reflectance refers to the surface albedo. The fact that the residual is very low in this view and lighting condition, where there is strong indirect illumination from the background, raises concerns. A strong deviation in reflectance suggests that under other viewpoints and lighting conditions, the error in rendering could be more significant, particularly in cases where the illumination does not produce such strong interreflections in the zebra's legs. To provide a clearer assessment, error maps, numerical metrics (e.g., PSNR), and perceptual metrics (e.g., SSIM) should be included in those specific cases to compare the ground truth (GT) with the results. Regardless, intrinsic decomposition still faces the well-known ambiguity between reflectance and shading, both local and global.

Please provide a clearer explanation of the choice of methods for comparison regarding the lighting setup. I understand that a method designed for a single light setting may not be able to handle

multiple light settings. However, a method that works with multiple random light settings should also be able to handle a single light setting. With that in mind, I wonder why methods like TensorIR, PIE-Net, and Ordinal are not used and compared in the Single setup. Could you please clarify this aspect?

It would be beneficial to include more examples of reflectance editing, relighting, and shading editing with multiple lights, along with visual comparisons to ground truth data and other methods. This is only done in Fig. 1, while the other figures focus on intrinsic image estimation but not on novel view renderings or virtual relighting. Novel view synthesis and relighting are reported in the tables as average metric values but lack visual examples. I believe that including this content, both in the paper and the supplementary material, is essential for a clearer understanding of the method's performance. Additionally, since the intrinsic decomposition of the proposed method appears to outperform current state-of-the-art solutions, these more reliable intrinsic images should contribute to better NVS and relighting. Showing this through visual examples would greatly strengthen the paper's contribution.

In Table 1, as stated in the caption, the best methods are marked in bold, and the second-best methods are underlined. This is true for reflectance, but there are some inconsistencies in the shading results. For example, in PSNR, both methods are in bold. In SSIM, the best method is neither bold nor underlined (i.e., 0.9225 - Ours Multiple). Similarly, for LPIPS, the best result (0.0798 - Ours Multiple) is shown in plain text. In the case of MSE, the second-best method is Ours Multiple (0.0055), not Ours Random (0.0072). A similar issue appears in Table 4 for PSNR, where no bold formatting is used.

In Fig. 4, I would recommend showing the real GT for shading for each case, rather than only reporting the GT for the Random Setting. This would make it easier to understand how the methods perform under different settings.

Presentation

In the abstract, the authors state that current methods for extracting intrinsic image components primarily focus on simple synthetic scenes and isolated objects. However, this claim is not entirely accurate. First, many recent methods handle real-world data effectively, including intrinsic image extraction in fairly complex indoor scenes. Second, this paper also focuses on isolated objects against relatively plain backgrounds. The authors should more clearly define the specific novelty they aim to address and provide a thorough discussion of how the existing literature falls short in addressing the problems under the given settings.

Since the code is not currently available, I would recommend changing the statement “The code and data are publicly available at https://github.com/liulisixin/MLI-NeRF” to “The code and data will be publicly available at https://github.com/liulisixin/MLI-NeRF.” Papers on Qeios can be updated frequently, so until the code is actually accessible, it would be better to avoid implying otherwise.

The sentence “...There are two related approaches to scene editing [8]: inverse rendering and intrinsic image decomposition...” is problematic. Inverse rendering and intrinsic image decomposition are not approaches to scene editing. Instead, editing is one of the tasks that can be performed after processing a dataset to extract intrinsic images or after using an inverse rendering pipeline to extract geometry and/or material properties. It is not accurate to define these techniques as "scene editing" methods. Additionally, reference [8] is a survey on intrinsic image decomposition, not on scene editing.

The equation numbers are not provided in the text, making it difficult to follow sentences like “...Since the Eq. (2) does not consider occlusion and other effects...”, as there is no Equation (2) indicated. The reader has to count the equations from the beginning to determine which one is being referred to. Similarly, in the sentence “...We use the equation $R = I \oslash S$ as a simplified version of Eq. (1) ignoring...”, it is unclear which equation is referred to as Equation (1). The same issue arises with the sentence “...Compared to the MLPcolor in Eq. (3)...”.

I would avoid sentences like “...we use high-fidelity pseudo intrinsic images to guide the intrinsic decomposition learning...”. A more accurate phrasing would be: “...we use the pseudo intrinsic images created in Step B and C to guide the intrinsic decomposition learning...”. There is no proof provided that these pseudo intrinsic images are of high fidelity.

In the results, I would clarify the discussion in Section 4.2 regarding grid and non-grid sampled light positions. For example, why are there more non-grid sampled lights than grid-sampled ones? Perhaps I’m missing some details, but the number of light positions in the grid-sampled case seems to depend on the grid step, whereas for non-grid sampled lights, it depends on the number of selected light positions. When the paper mentions “...handle hundreds of light positions...,” does this refer to handling them all at once? Does the use of non-grid lights imply that lighting conditions can be approximated by a multiple-light setup (e.g., four lights together), where each light corresponds to a specific position? Please provide a more detailed explanation of this point.

Fig. 5(b) is misleading. It doesn't make much sense to compare the original images (as GT) with the reflectances. At the very least, I would not refer to them as GT images. If the goal is to show how the

method removes the effects of shading by comparing an original image with the reflectance signal, I would suggest replacing the label "GT" with "Original Image" for clarity.

The authors mentioned that they “...have also submitted a supplementary video to showcase the results of our method on all datasets....”. However, I could not find it on the submission page. Please clarify where the video can be accessed.

References

Please adjust the format of the references, as the extensive use of italics makes them difficult to read.

Regarding reference [3], I would not directly associate it with NeRF as suggested in the first sentence of the introduction. In fact, it is somewhat complementary to NeRF. While NeRF computes volume density, the paper in question opts for neural surface reconstruction to provide "...more direct modeling of surfaces...for photogrammetric surface reconstruction problems...". The authors of that paper note that "...A problem of NeRF and its variants [1,30,43,46], however, is the question of how an isosurface of the volume density could be defined to represent the underlying 3D geometry...". While NeRF focuses on extracting geometry and color for novel view synthesis, Neuralangelo serves a different purpose—"...scene representations with better-defined 3D surfaces...". Given the vast landscape of the NeRF literature, I would suggest avoiding citing papers that are only marginally related to NeRF.

Please avoid duplicated references ⇒ [17] and [21]

- Li Z, Snavely N. "Learning Intrinsic Image Decomposition from Watching the World." In:Computer Vision and Pattern Recognition (CVPR); 2018.
- Li Z, Snavely N. "Learning intrinsic image decomposition from watching the world". In:Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9039–9048

When possible, please avoid using the arXiv citation and instead use the appropriate published citation:

- Ling, Jingwang, Zhibo Wang, and Feng Xu. "Shadowneus: Neural sdf reconstruction by shadow ray supervision." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 175-185. 2023.

- El Helou, Majed, Ruofan Zhou, Sabine Süsstrunk, Radu Timofte, Mahmoud Afifi, Michael S. Brown, Kele Xu et al. "AIM 2020: Scene relighting and illumination estimation challenge." In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 499-518. Springer International Publishing, 2020.
- Puthussery, Densen, Hrishikesh Panikkasseril Sethumadhavan, Melvin Kuriakose, and Jiji Charangatt Victor. "Wdrn: A wavelet decomposed relightnet for image relighting." In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 519-534. Springer International Publishing, 2020.
- Wang, Li-Wen, Wan-Chi Siu, Zhi-Song Liu, Chu-Tak Li, and Daniel PK Lun. "Deep relighting networks for image light source manipulation." In Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 550-567. Springer International Publishing, 2020.

Please ensure consistency in the citation format, especially when referencing the same journal. For example, the following citations come from the same journal but are formatted quite differently.

- ACM ToG
  - Sun T, Barron JT, Tsai YT, Xu Z, Yu X, Fyffe G, et al. Single image portrait relighting.ACM Trans. Graph. 38 (4): 79--1, 2019
  - Pandey R, Orts-Escolano S, Legendre C, Haene C, Bouaziz S, Rhemann C, Debevec PE, Fanello SR (2021). "Total relighting: learning to relight portraits for background replacement." ACM Trans. Graph.. 40: 43--1
- CVPR
  - Toschi M, De Matteo R, Spezialetti R, De Gregorio D, Di Stefano L, Salti S. "ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 20762-20772
  - Li Z, M\u00fcller T, Evans A, Taylor RH, Unberath M, Liu M-Y, Lin C-H. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2023
  - Zhang Y, Sun J, He X, Fu H, Jia R, Zhou X (2022). "Modeling Indirect Illumination for Inverse Rendering". In: CVPR, 2022.

    - Li Z, Snavely N. "Learning Intrinsic Image Decomposition from Watching the World." In:Computer Vision and Pattern Recognition (CVPR); 2018.
    - Zhang K, Luan F, Wang Q, Bala K, Snavely N. "PhySG: Inverse Rendering with Spherical Gaussians for Physicsbased Material Editing and Relighting." In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021.
- For example, CVPR is cited in six different ways, such as “IEEE conference on computer vision and pattern recognition”, “Proceedings of the IEEE conference on computer vision and pattern recognition”, “Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition “, “CVPR”, “Computer Vision and Pattern Recognition (CVPR)”, “The IEEE/CVF Conference on Computer Vision and Pattern Recognition”

English / Typos

Section 4.6 ⇒ “...In the all lights setting...” should be “...In the all light setting...”. Here, "all light" functions as an adjective, so it does not require the plural form. It's similar to the expression "a two-bulb illumination".

References ⇒ “Li Z, M\u00fcller T, Evans A, Taylor RH, Unberath M...” Be careful and replace “\u00” with the proper character

Declarations

Potential competing interests: No potential competing interests to declare.