Rapid Estimation of Optical Properties for Simulation-Based Evaluation of Pose Estimation Performance

Iversen, Thorbjørn Mosekjær; Wilm, Jakob; Kraft, Dirk

Published in:

DOI:
10.1109/IROS40897.2019.8967553

Publication date:
2019

Document version
Accepted manuscript

Citation for published version (APA):
Rapid Estimation of Optical Properties for Simulation-Based Evaluation of Pose Estimation Performance

Thorbjørn Mosekjær Iversen, Jakob Wilm and Dirk Kraft

Abstract—A growing trend in computer vision is the use of synthetic images for the evaluation of computer vision algorithms such as 3D pose estimation. This is partly due to the availability of high-quality render engines, which provide highly realistic synthetic images. However, the realism of the rendered images, and thus the reliability of the evaluations, strongly depends on how accurately the scenes are modeled and it requires considerable time and knowledge to do the modeling manually. Automating the modeling process is therefore crucial for making the rendering of photo-realistic synthetic images accessible to the wider robotics community.

We present a method for automatically modeling object and light properties for rigid, opaque plastic objects commonly found in industry. Our method relies on recordings of the environment captured with a consumer 360° camera to model the light, and on analysis-by-synthesis to estimate the optical properties of the objects. We show that the synthetic images rendered based on our automatic modeling method can be used to predict the overall performance of a monocular 3D pose estimation algorithm.

I. INTRODUCTION

The evaluation and tuning of computer vision algorithms rely on visual input. Depending on the context, it can be time-consuming, impractical, or infeasible to capture the needed vision data. Due to the increase in the computational power of modern PCs as well as the improvement of render engines, it has become increasingly popular to rely on synthetic images in addition to real ones [1].

The generation of realistic synthetic images requires accurate models of the light and objects in the scene. Manual modeling takes considerable effort, which leads to a trade-off between model accuracy and time invested in the modeling process. The computer graphics community has a strong focus on increasing the realism of rendered images and consequently provides sophisticated methods and specialized equipment, such as gonioreflectometers, for accurately measuring the visual properties of objects. However, many of these methods prioritize measurement accuracy over measurement speed or rely on expensive specialized equipment. As a result, these methods are seldom used in the computer vision and robotics communities where synthetic images are mainly used as a means to an end.

We propose a method for automating the modeling of light and optical object properties for opaque plastic objects. We use recordings of the environment captured with a consumer 360° camera to model the illumination and analysis-by-synthesis to estimate the optical properties. This allows for the generation of highly realistic synthetic images with a minimum of modeling required. Our method thus provides a compromise between modeling speed and image realism. While the level of realism is less than what can be expected if rigid measurements of the bi-directional reflectance distribution function (BRDF) was performed, the method is simple, fast and still results in highly realistic images (see Fig. 1). We evaluate our method by evaluating the performance of a monocular 3D pose estimation algorithm using both real images and synthetic images modeled using our method.

This paper is structured as follows. First Section II presents an overview of the related literature. Section III then presents our method for modeling the light and optical object properties using analysis-by-synthesis. Section IV describes the details of the evaluation, and Section V presents the results of the evaluation and a discussion on the found results. Finally, Section VI concludes on our method and results.

II. RELATED WORK

In the computer vision literature, synthetic images are utilized in a wide range of use cases. They are for instance used in the training of algorithms [2], [3], [4], analysis of uncertainties [5], evaluation of algorithms [1], [6], [7], analysis of visual features [8], [9] and offline programming and prototyping [10], [11]. The methods used for generating the images, the realism of the images and the methods of evaluation differs greatly depending on the use case. In a study of the use of graphics simulations for the evaluation of vision algorithms, Veeravasarapu et al. [12] found that it is still an open question what level of realism is needed.
to perform meaningful evaluations. It is, therefore, crucial to include a dedicated evaluation of the realism of synthetic images when these are used for evaluation.

In [13], Rossmann et al. present work that explicitly evaluates the realism of synthetic images. They present a virtual testbed, which provides a framework for simulating various vision sensors, for use in the development and initial testing stage of a robotic system. Their system relies on OpenGL rasterization with a subsequent application of filters. This allows them to render synthetic images in real time.

One of the disadvantages of using rasterization is that some of the subtle interactions between light and material are lost. In order to render images with a high degree of realism, Medeiros et al. [1] use a physically based rendering pipeline to simulate a structured light scanner. The synthetic scanner is then used to make a quantitative evaluation of different structured light scanning techniques, taking effects such as subsurface scattering and projector defocus into account.

In [7], Irgenfried et al. explicitly evaluates the realism of synthetic images based on how a given computer vision algorithm performs when applied to a real image and a corresponding synthetic image. In their work, the performance of different segmentation algorithms is evaluated on both synthetic and real images. They report similar performance when a segmentation algorithm is applied to simple scenes. However, for complex scenes, there are some discrepancies between the segmentation performed on the real and synthetic images. They attribute the discrepancies to insufficient modeling of the global illumination.

The above works require that the scene has been modeled. This includes lights, object material, and object and scene geometry. Since state-of-the-art rendering engines are able to reproduce almost any kind of real-world effect, the differences between synthetic and real images can be almost entirely attributed to insufficient modeling. A common method for capturing the BRDF of a surface is to use a gonioreflectometer. Unfortunately, such a device is both expensive to acquire and time-consuming to use. For this reason, the development of more rapid methods is an active research field (see e.g. the work by Mukaigawa et al. [14]).

An alternative to traditional measurements of object properties is analysis-by-synthesis. This method compares a rendering to a real image and iteratively adjusts the model parameters until the reconstruction error has been minimized. The model parameters at the minimum are then estimates of the corresponding real-world parameters. An example of this is in the work by Corso et al. [15] where analysis-by-synthesis is used to infer the concentration of apple particles in a glass of apple juice based on an image. Another example is the work by Stets et al. [16] where analysis-by-synthesis is used to determine the absorption-coefficient of glass.

Our work is strongly inspired by Stets et al. [16]. In their work, a scene, consisting of glass objects on a white tablecloth, is reconstructed by fusing multi-model sensor data. Using highly accurate, specialized calibration methods, they are able to render highly realistic synthetic images which are comparable to corresponding real images on a pixel-to-pixel level.

Our work uses a streamlined image creation pipeline, which relies on analysis-by-synthesis, to infer the object parameters and on recordings of the environment to model the illumination. Although our method does not achieve the same accuracy as [16], it does not rely on expensive specialized equipment and time-consuming recordings to model the scene. This makes our method very useful for many robotics applications, in that the performance of vision equipment can be realistically simulated and optimized before investing engineering effort in implementing a physical cell.

We show that the images rendered based on modeling using our method are comparable to what is achieved by manual modeling. Additionally, we show the images are sufficiently realistic to predict the overall performance of the pose estimation algorithm described in [17]. In this work, pose estimation will refer to the use of this method. To the best of our knowledge, our work is the first to use automatic object material and light modeling to facilitate fast and easy simulation-based evaluation of pose estimation performance.

III. Method

In order to render a synthetic image, the scene must be described by a set of models which define relevant entities such as the light, the poses of the objects, and the geometry and material properties of the objects. The choice of models dictates the theoretical maximum limit of realism achievable in the image. This section first describes and discusses the models used in our method which are all commonly used models. This is followed by a description of our method for automatic estimation of optical object properties.

A. The models

The scene model is split into parts which are approached separately. The three parts are the lighting, the geometry, and the optical properties of the object. The following describes each part in more detail.

Lighting. Ideally, all light sources in the scene should be modeled by their position, intensity, and distribution of light emission. While this is possible in a controlled lab environment, industrial scenes generally include many light sources. As light also reflects off objects which are not visible in the image, a complete light model would require the entire scene to be modeled including objects outside the field-of-view. To avoid this, the light is modeled by an environment map: the scene is modeled as consisting of local objects illuminated by an environment which is infinitely far away. This approximation allows the light to bounce between local objects in the scene without interacting with the environment.

Modeling the light as coming from infinitely far away can potentially introduce a significant parallax error. It is possible to correct for this using parallax correction. However, this would require a manual estimate of the position of the light in the environment. As our method has been developed to automate the modeling process, we have
Scene and object geometry. In this work, object geometry refers to both the shape of the object as well as the relative 6D poses of objects, light, camera, and, in the case of the experiments, markers. Our method has been developed for industrial cases, which means that CAD models are available for all objects of interest. When this is not the case it is possible to either reverse engineer or 3D scan the objects. It is therefore assumed that digital representations of object shapes are available as triangle meshes.

Optical properties of the object. The appearance of an object depends on the incoming light and how this light interacts with the material. While it is impossible to take into account all physical effects taking place, the computer graphics literature provides models which approximate almost any aspect of light interaction. Our study involves opaque plastic objects since these are common in industry and can be fairly accurately described by a limited set of render parameters.

We use Pixar Renderman v21 as our rendering engine. The object material is modeled using the Renderman material PxrSurface. Based on initial experiments and an analysis of the available model parameters, we conclude that opaque plastic objects can be relatively accurately modeled using only three parameters: color, surface roughness, and refractive index. Note that this model treats surface roughness as a material parameter regardless of the fact that surface roughness is the microscopic structure of the object geometry. As a result, objects which have surface areas of different roughness are modeled as consisting of different materials.

B. Estimation of model parameters

The achieved realism of the synthetic images largely depends on the model parameters. The following describes our method for estimating these parameters (the lighting, the geometry, and the optical properties of the object). The method consists of three steps. In the first step, images of the environment are recorded using a Theta V 360° camera and merged into a high dynamic range (HDR) environment map which contains the relative light intensity of the environment. In the second step, an image of the object of interest is captured under specialized calibration conditions. In the third step, the light intensity and object material parameters are optimized jointly. A data flow diagram illustrating our method is shown in Fig. 2.

Recording of HDR environment map. Images of the environment are recorded using a Ricoh Theta V camera, which is a consumer camera that can capture 360° images which cover the full azimuth angle range ($\phi \in [0^\circ, 360^\circ]$) and elevation angle range ($\theta \in [-90, 90]$). These images are captured by two image sensors with corresponding fisheye lenses positioned such that one covers the front-facing hemisphere and the other covers the back-facing hemisphere. Immediately after capture, the two fisheye images are stitched where the hemisphere borders meet and mapped to and saved as a single equirectangular image as shown in Fig. 3b.

In order to perform seamless stitching, the Theta V distorts the two input images before stitching. This distortion is strongest close to the image border. Initial experiments have shown that strong light sources close to the image border tend to be attenuated. As a result, this part of the image is unsuitable for reliable recording of relative light intensity which is needed for the creation of an HDR environment map. To mitigate the effect, we place the camera with the front lens pointing upwards and the back lens covered. This results in images where the distorted regions are strongest close to the table horizon. This approach could potentially lead to inaccurate environment maps if there are strong light sources at a low elevation angle above the table. However, in many industrial applications, light sources are placed mainly above the table surface. This means that the distortion at the edges has a relatively low practical impact on the rendered images. This workaround makes it possible to use the Theta V without the need for low-level hardware access.

A number of images of the environment are captured at varying exposures. The images are merged into an HDR image usingDebevec’s method [18]. We use OpenCV's implementation which generates the images in two steps. First, the response curve of the camera is calibrated, after which the images, which are taken at different exposures, are merged to a single HDR environment map. To limit the negative impact, which image stitching has on the calibration, only a subregion in the environment images is used for the calibration of the response curve. The region is the central 25% of the front camera illustrated in Fig. 3. After the
response curve has been calibrated, the environment images are merged into an HDR environment map using the full image region. The rotation of the environment map relative to the workspace is found by pose estimating the Theta V. The correspondence between the coordinate frame of the Theta V and the environment map is illustrated in Fig. 3. This process captures all model parameters relevant to the light except the absolute intensity of the light, which is included in the analysis-by-synthesis optimization discussed later in this section.

**Recording of the analysis-by-synthesis image.** The optical properties of the object, which are color, refraction index, and surface roughness, are determined through minimization of the reconstruction error. This minimization is based on a single real image. It is therefore important that this image contains the information necessary for the minimization. Practically, this means that the image must be captured from a viewpoint where both directly reflected light and diffusely reflected light is visible. Furthermore, if the object contains surfaces of a different color or surface roughness, the image should either have all these surfaces in view, or several images should be captured such that each surface can be treated separately in separate images.

The most important visual clues for the determination of surface roughness and refractive index are the specular highlights. It is therefore important to keep these as clear and undisturbed from indirect light as possible. The analysis-by-synthesis image is therefore recorded inside a box shielded from outside light using black curtains. Furthermore, the table is covered with black low-reflection flocked paper (Thorlabs BFP1) to provide a black background. The only light source is a high power LED diode, strong enough to illuminate the scene. It is important to note that these unique lighting conditions, especially the use of flocked paper, provide a number of benefits. Since very little light is reflected by the table, the material properties of the table do not have to be modeled and the table surface does not have to be pose estimated. Furthermore, the perfectly black diffuse background increases the rendering speed of the synthetic images, since no light is reflected by the surface, which results in fewer light bounces for the light rays. In our setup, the black surface reduced the rendering time to less than 10% of the original rendering time (from approximately 570 seconds to 55 seconds\(^1\)). The setup used to capture the analysis-by-synthesis image is shown in Fig. 4a. An example of an image captured under the described illumination conditions is shown in Fig. 5a.

**Optimization of object parameters.** In each iteration of the analysis-by-synthesis optimization, the captured image is compared to a corresponding synthetic image. In order to render such an image, the pose of the object must be estimated. In many cases, this can simply be done using the monocular pose estimation algorithm in [17]. However, if this leads to a poor pose estimate, it might be necessary to change the illumination or pose estimate the object in different images and then use markers to define a common frame of reference (similar to how the ground truth pose is estimated in Section IV). Once the pose of the object has been estimated, all relevant model parameters that are not determined in the optimization have been defined.

The optimization is performed as a joint optimization over the object material model parameters and the light intensity. Since only the relative light intensity is contained in the HDR environment map and the color of the object is unknown, the captured image does not contain enough information to determine whether the image shows a light object under low illumination or a dark object under strong illumination. This means that the color determined by the optimization does not necessarily reflect the object color in absolute terms. However, in the context of this work, this is acceptable provided that the modeled parameters lead to rendered images similar to real images. Note that once the light and object parameters have been determined, the light intensity must be held constant in subsequent optimizations.

\(^1\)The images were rendered on an Intel Core i7-4600U CPU
since a scene can contain several different materials but only one value for light intensity. Additionally, if the material properties of an object have already been determined in a previous experiment and new environment light has been recorded, the analysis-by-synthesis optimization can be used to only optimize the light intensity.

The reconstruction loss function used is the sum of absolute differences (SAD) between pixels in the real and synthetic image. The method assumes that the estimated pose of the object is accurate enough that the pixels can be compared on a one-to-one basis. Although initial experiments showed that this level of accuracy is rarely achieved, they also showed that the pose is rarely displaced by more than a few pixels. For most pixels, except edge pixels, the intensity doesn’t change much with a few pixels displacement. Consequently, the SAD cost function performs well. As only pixels corresponding to the object are of interest, a mask indicating foreground pixels in the synthetic image is defined.

The analysis-by-synthesis optimization can therefore be stated as follows:

$$
\mathbf{x}' = \arg \min_{\mathbf{x}} \sum_i M_i | R_i - S_i(\mathbf{x}) | \tag{1}
$$

Here, $M_i$, $R_i$, and $S_i$ refer to the $i$th pixel value of the mask, real image, and synthetic image respectively. Furthermore, $\mathbf{x}$ refers to a vector of the model parameters which are optimized (light intensity, surface color, surface roughness, and refractive index). $\mathbf{x}'$ is used to denote the final model parameters.

The optimization uses the Bound Optimization by Quadratic Approximation (BOBYQA) algorithm [19]. The algorithm is a trust region based derivative-free constrained optimization, which excels at optimizing black-box functions. Each parameter is multiplied by a weight such that the parameters are of approximately equal magnitude. The starting point of the optimization should be a reasonable guess of the parameter values of a generic plastic object. Furthermore, the initial guess of surface roughness should be set such that any highlights visible in the real image are also rendered in the initial synthetic image. Otherwise, the optimization is likely to not find the global optimum. The synthetic image corresponding to Fig. 5a rendered based on the final model parameters is shown in Fig. 5b.

IV. Evaluation experiment

The evaluation consists of three experiments which investigate the realism of synthetic images rendered under different scene conditions. The following section first presents the method of comparison, followed by an overview of the three experiments. Finally, the details of the experimental setup and the evaluation procedure are presented.

A. Algorithm-realism

Before the evaluations can be performed, a measure of realism must be defined. Our method has been developed for fast and easy rendering of highly realistic synthetic images with a focus on evaluating pose estimation performance. It is therefore not the appearance of an image that is important, but rather how the pose estimation algorithm performs on the image. Therefore, the measure of realism is the similarity between the output of the pose estimation algorithm when applied to a real image and corresponding synthetic image. We introduce the term algorithm-realism to clarify that we are interested in predicting algorithm performance, and not visual realism as perceived by human observers.

In this evaluation, the algorithm-realism is determined for the monocular 3D pose estimation algorithm described in [17]². This pose estimation algorithm relies on the detection of gradients which are predominant features used in many pose estimation techniques. The algorithm-realism is assessed by applying the pose estimation algorithm to a real and corresponding synthetic image and comparing three output metrics of the pose estimation algorithm: the similarity score of the pose estimate, the magnitude of the translational part of the pose error, and the magnitude of the rotational part of the pose error.

The similarity score is chosen as a measure of algorithm-realism since it is an essential part of the pose estimation algorithm where it is used to rate competing pose candidates. The pose estimate leading to the highest similarity score is returned as the most likely estimate. The similarity score is a measure of similarity between the gradients of $n$ edge pixels in a pose-dependent 2D model ($s_i$) and the gradients in the image ($m_i$), and it is defined as [17]:

$$
c = \frac{1}{n} \sum_{i=1}^{n} \frac{\langle m_i, s_i \rangle}{||m_i|| \cdot ||s_i||} \tag{2}
$$

B. Overview of experiments

Each experiment in the evaluation investigates a different aspect of our method. The first experiment investigates if the modeling performed with our method leads to a similar level of algorithm-realism as what can be achieved using manual modeling. The second experiment investigates the algorithm-realism when the object properties are transferred to a different object of similar material. The third experiment investigates the algorithm-realism for a scene with complex lighting. An overview of the experiments and the synthetic images rendered in each experiment is provided in Table I.

‘Object’ refers to the object in the scene which is either a cap or a socket (see Fig. 6). ‘Light model’ refers to how the light is modeled in the synthetic images. The illumination in experiment A and B is a high powered LED diode (Fig. 4a) while the illumination in experiment C is the light from the lab (Fig. 4b). The environment light model refers to the model based on recordings of the environment described in Section III. Point light refers to modeling of the light as a point light positioned at the camera center (see e.g. Fig. 5c).

‘Material’ refers to the values of the object properties. Manual means that the values have been chosen manually,

²The implementation find_shape_model_3D from the MVTec Halcon 13 library is used
AbS means that our analysis-by-synthesis method has been used, and from A3 means that the values determined from analysis-by-synthesis in experiment A3 are used.

Note that the reason for rendering synthetic images based on a point light model is to evaluate the potential improvement in algorithm-realism from using a realistic light model.

C. Experimental setup

In each experiment, real images are captured at various viewpoints surrounding the object. While the exact viewpoints differ between the three experiments, each of the three recorded datasets contains images recorded at four different elevation angles: the first image is a top-down view while the rest are recorded at three different elevation angles in approximately even steps of azimuth angle, such that the set of viewpoints in each experiment approximately covers a hemisphere around the object. The distance to the object is approximately 45-55cm.

The setup used to record the datasets for the three experiments is shown in Fig. 4. A monochrome camera (Basler acA1300-60gm with a Spacecom JHF8MK lens) is mounted on a flexible photo-arm. The setup contains a strong LED diode light (Thorlabs MCWHLP1). Furthermore, the top of the box and the curtains can be removed to allow regular lab light to illuminate the setup. To provide a common reference for all camera views, three ChArUco calibration markers [20] are positioned around the workspace center (where the objects are placed) and their relative poses are calibrated. The ChArUco markers are slightly tilted such that they can be viewed from a low elevation angle.

To evaluate the performance of the pose estimation algorithm, a good estimate of the ground truth object pose is required. This pose is estimated by first pose estimating the object in each image. Each pose estimate is then manually labeled as either ‘correct’ or ‘incorrect’. The ground truth pose \((R_{gt}, t_{gt})\) is estimated as the mean of the ‘correct’ pose estimates (translation is the mean of the translation vectors and rotation is the mean of the angle-axis rotation vectors).

The environment maps were recorded with the Theta V camera in the workspace center. Because of the high power of the diode, the shortest shutter time for the Theta V camera, \(1/25000\) s, was not short enough to avoid pixel saturation. To circumvent this, the environment images were recorded with the diode power turned down to approximately 4% of its full strength. This is only possible because the environment map contains the relative light intensity and the diode is the only light source in experiment A and B. The environment was recorded with exposures from 15 seconds to \(1/25000\) seconds in 29 steps. Pixel saturation was not an issue with the lab environment light, so these images were recorded with exposures from 1 second to \(1/25000\) seconds in 17 steps. The pose of the Theta V was estimated from a single image containing both the Theta V and the ChArUco markers.

Using the above calibration it is possible to render synthetic images corresponding to the real images in the datasets. The full method for estimating the light and optical object properties is used in A1 to determine the optical properties of the cap. The cap is made of a single material and has two surface areas of different roughness: a shiny and a matte. This is illustrated in Fig. 6a. The analysis-by-synthesis optimization was done simultaneously for the two surfaces, thus leading to a joint optimization over light intensity, object color, shiny surface roughness, matte surface roughness, and refraction index.

In each image, real or synthetic, the object is pose estimated. The similarity score, \(c_i\), is provided by the pose estimation algorithm, and the magnitude of the translational and rotational part of the pose error, \(\epsilon_t\) and \(\epsilon_r\), respectively, can be computed using the following:

\[
\epsilon_t = \| R^{\top}_{gt}(t_{est} - t_{gt}) \| \\
\epsilon_r = ||\mathbf{r}||, \]

where \(\mathbf{r}\) is the angle-axis vector corresponding to the rotation matrix \(R^{\top}_{gt}R_{est}\). Here, \(t\) and \(R\) denotes the translation vector and rotation matrix part of a homogeneous transformation matrix. The subscript ‘est’ refers to the pose estimate in the image and ‘gt’ refers to the ground truth pose. The algorithm-realism is determined by plotting \(c_i\), \(\epsilon_t\), and \(\epsilon_r\) as functions of viewpoint for both the real and synthetic images.

Both the cap and the socket are challenging objects to pose estimate since images captured for 90-degree rotations around their z-axes look almost identical. In [17] it is pointed out that the algorithm is not able to distinguish between views where the appearance of the object, as seen from different directions, are identical. This difficulty, coupled with the fact that small differences between the real and synthetic image are practically unavoidable, means that it cannot be predicted from the synthetic images which of the four rotations will be estimated in the real images. In order to still perform a meaningful evaluation, we therefore treat \({0^\circ, 90^\circ, 180^\circ, 270^\circ}\)-rotations as equivalent.
Fig. 7 shows the similarity score \( c \), \( \epsilon_t \), and \( \epsilon_r \) as functions of viewpoint index for the experiments described in Table I.

The results for the first experiment show that there is no difference between manually modeling the object parameters (\( A2 \)) and using our analysis-by-synthesis method (\( A3 \)) as functions of viewpoint index for the experiments described in Table I. The difference is negligible. This is an important result since it means that the time-consuming manual modeling process can be replaced with our method with no loss in algorithm-realism. Furthermore, the experiment shows that the object properties of the synthetic images based on modeling using our method (\( A3 \)) are relatively high. This means that the synthetic images of our method enable the prediction of the overall pose estimation performance. The first experiment also shows that the algorithm-realism of the synthetic images which does not use environment light (\( A1 \)) is lower than when environment light is used. It is observed the images using the environment light model, generally, predicts the pose estimation performance to be too good compared to reality.

The second experiment shows that it is possible to transfer the optical object properties of the cap to the socket. The algorithm-realism of the new synthetic images (\( B2 \)) is comparable to the images the first experiment (\( A3 \)). Furthermore, the second experiment shows that the synthetic images which do not use environment light (\( B1 \)) greatly overestimate the pose estimation performance. This shows the importance of using a realistic model of illumination.

The third experiment investigates how well the model parameters transfer to complex lighting. By comparing the results of experiment A and C it can be observed that the lighting in the robotics lab leads to better pose estimation performance than when the LED diode is the only illumination. With a few exceptions, this is correctly predicted from the synthetic images using the environment light model (\( C2 \)). This is also predicted from the images using the point light model (\( C1 \)). However, previous experiments showed that images based on the point light model tend to overestimate performance (high recall, low precision), so correct predictions of good performance is to be expected.

Based on the three experiments it can be concluded that the synthetic images based on our method are sufficiently realistic to predict the overall pose estimation performance of [17]. It has also been shown that once the model parameters have been determined they allow for the generation of algorithm-realistic images of different objects with similar material properties and scenes with different illumination.
While our method is able to predict the overall pose estimation performance, it has not been possible to make detailed predictions of the pose uncertainty. A possible explanation is that the current setup has a position uncertainty of a few millimeters. A few millimeters or degrees of calibration error can have a significant impact on how shadows fall on the object. The pose estimation algorithm defines the best pose estimate as the one leading to the highest similarity score. Initial experiments for the cap have shown that there are often different competing poses with almost the same similarity score. Consequently, even small differences in object appearance can lead to very different poses estimates, even if the similarity scores are almost identical. It is left for future work to investigate if improved calibration and analysis of competing poses with high similarity score will allow for detailed predictions of pose uncertainty.

VI. CONCLUSION

In this work, a method for automated modeling of light and optical object properties has been presented. The light is recorded using images of the environment captured with a consumer 360° camera and the optical properties of an object are determined by using analysis-by-synthesis. The method has been used to generate highly realistic synthetic images with which the overall performance of a monocular pose estimation algorithm has been evaluated in simulation.

The evaluation of the realism of the generated synthetic images has been done in the context of predicting pose estimation performance. Since the synthetic images are rendered with the purpose of evaluating a pose estimation algorithm in simulation, it has been argued that the measure of realism should reflect this. The evaluation, therefore, applied the pose estimation algorithm to both real and synthetic images and compared the results. This comparison of algorithm output on real and synthetic images has been referred to as the algorithm-realism of the synthetic images. The evaluation showed that our analysis-by-synthesis method leads to images with an algorithm-realism comparable to images where manual modeling has been used. As a result, our automated method can replace time-consuming manual modeling within the problem domain. It was also shown that the generated synthetic images were sufficiently realistic to predict the overall performance of the pose estimation algorithm. Our method was shown to be applicable to different objects and lighting conditions.

Future work should investigate how the algorithm-realism of the generated images can be improved. One approach would be to include a measure of algorithm-realism in the cost function in the analysis-by-synthesis optimization, such that the images are explicitly optimized for this. Another possibility is to include fine-tuning of the object pose in the optimization. This could reduce the discrepancy between the real and the rendered image leading to an improved analysis-by-synthesis optimization. Finally, our method should be expanded to include more complex materials.

References