Universität des Saarlandes

Fachrichtung 6.1 – Mathematik

Preprint Nr. 355 (revised)

Physically Inspired Depth–from–Defocus

Nico Persch, Christopher Schroers,
Simon Setzer and Joachim Weickert

Saarbrücken 2016

Fachrichtung 6.1 – Mathematik
Universität des Saarlandes

Preprint No. 355 (revised)
submitted: June 23, 2016

Physically Inspired Depth–from–Defocus

Nico Persch
Mathematical Image Analysis Group, Dept. of Mathematics and Computer Science, Campus E1.7, Saarland University, 66123 Saarbrücken, Germany
persch@mia.uni-saarland.de

Christopher Schroers
Mathematical Image Analysis Group, Dept. of Mathematics and Computer Science, Campus E1.7, Saarland University, 66123 Saarbrücken, Germany
schroers@mia.uni-saarland.de

Simon Setzer
Mathematical Image Analysis Group, Dept. of Mathematics and Computer Science, Campus E1.7, Saarland University, 66123 Saarbrücken, Germany
setzer@mia.uni-saarland.de

Joachim Weickert
Mathematical Image Analysis Group, Dept. of Mathematics and Computer Science, Campus E1.7, Saarland University, 66123 Saarbrücken, Germany
weickert@mia.uni-saarland.de

Edited by
FR 6.1 – Mathematik
Universität des Saarlandes
Postfach 15 11 50
66041 Saarbrücken
Germany

Fax: + 49 681 302 4443
e-Mail: preprint@math.uni-sb.de
WWW: http://www.math.uni-sb.de/

Abstract

We propose a novel variational approach to the depth–from–defocus problem. The quality of such methods strongly depends on the modelling of the image formation (forward operator) that connects depth with out–of–focus blur. Therefore, we discuss different image formation models and design a forward operator that preserves essential physical properties such as a maximum–minimum principle for the intensity values. This allows us to approximate the thin–lens camera model in a better way than previous approaches. Our forward operator refrains from any equifocal assumptions and fits well into a variational framework. Additionally, we extend our model to the multi–channel case and show the benefits of a robustification. To cope with noisy input data, we embed our method in a joint depth–from–defocus and denoising approach. For the minimisation of our energy functional, we show the advantages of a multiplicative Euler–Lagrange formalism in two aspects: First, it constrains our solution to the plausible positive range. Second, we are able to develop a semi–implicit gradient descent scheme with a higher stability range. While synthetic experiments confirm the achieved improvements, experiments on real data illustrate the applicability of the overall method.

1 Introduction

Depth–from–defocus methods are designed to handle the limited *depth–of–field (DOF)* of optical systems. This limited DOF is caused by the fact that a lens can only focus points at a certain distance that is given by the *focal plane*. Depth–of–field denotes the distance range in which objects still appear acceptably sharp. Points displaced from the focal plane are imaged blurred, where the out–of–focus blur increases as their offset becomes larger. Both the position of the focal plane and the size of the depth–of–field depend on the optical settings of the system. Macro photography and microscope imaging are typical examples of imaging systems suffering from a very limited depth–of–field.

To imitate an acquisition as it would be done by such a limited DOF imaging system, computer graphics methods can simulate the local blur on the basis of the local depth information. This is called *depth–of–field simulation.* In principle, depth–from–defocus is the inverse operation to the depth–of–field simulation: Its goal is to infer the depth map from the local blur. Moreover, as a by-product, one is able to compute the completely sharp image that would result from an infinite depth–of–field. Since one cannot distinguish between a blurred texture and out–of–focus blur, several images of the same

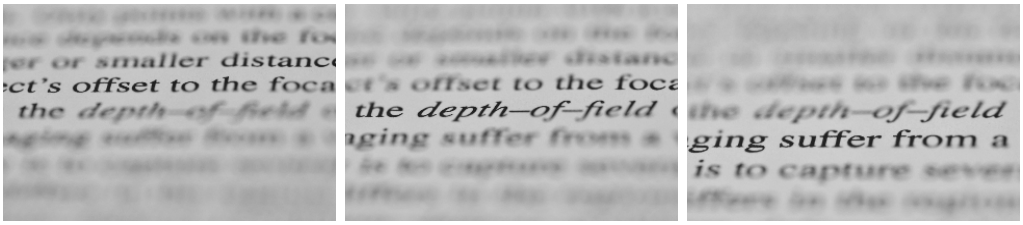

Figure 1: Focal stack. Each slice captures the same static scene but differs in their focal settings. The gradual transition from sharp to blur within each image corresponds to the depth profile.

static scene but varying in their focal plane are necessary. Then each of these images contains different regions that are projected sharply (cf. Figure 1).
While the forward problem of simulating the depth–of–field effect always has a unique solution, depth–from–defocus is a typical inverse problem that is ill-posed: For instance, blurring a homogeneous region creates a focal stack that does not allow to reconstruct a unique depth map. Ill–posed problems can be addressed by embedding them in a variational framework. This allows to extract a unique solution as the minimiser of some energy functional that involves an additional smoothness assumption [4]. To keep the numerical complexity reasonable, many variational models for the depth–from–defocus problem involve only a relatively small set of simplified assumptions, e.g. requiring local equifocality. As a consequence, it can happen that their solution reflects the physical reality only to a limited extent. Approaches that do not rely on an equifocal assumption at all, can also violate important physical principles such as a maximum–minimum principle for the intensity values. The goal of our paper is to propose a variational model for the depth–from–defocus problem that comes closer to the physical reality. Before going into more details, we discuss the relevant literature first.

Related Work. Pentland is one of the pioneers in the field of depth–from–defocus. He suggested to estimate the depth using defocus information by considering the blurriness of sharp edges or patches in different recordings [28]. Furthermore, he discussed the use of a Gaussian as a suitable point spread function (PSF) w.r.t. diffraction effects.
While Pentland assumed one image to be in focus as reference, Subbarao [35] refrained from such a restriction. Continuing his work in [36], he suggested a depth–from–defocus approach using a convolution/deconvolution transform in the spatial-domain. The sensitivity to noise of the latter approach is investigated in [37].

Bailey et al. [2] determined the level of blurriness using the method of [17] followed by a normalised convolution. Then they use the relation between depth, position of the focal plane, and the blurriness to recover the depth.
In [24], Namboodiri and Chaudhuri exploited the equivalence between Gaussian blurring and linear diffusion to simulate the depth–of–field effect. While this assumes the scene to have a constant depth, in [14, 25, 26, 41] the idea is extended by considering nonlinear isotropic diffusion that allows spatial changes of the diffusivity corresponding to the depth profile. Involving also directional information, anisotropic diffusion is used in [12, 16] to solve the depth–from–defocus problem.
Markov random fields can also be employed to handle the depth–from–defocus problem [10, 5, 26]. Among those approaches, Bhasin and Chaudhuri [5], investigate how the PSF has to be iteratively corrected at strong depth changes, where partial occlusions occur. However, their work is limited to only two focal planes (two defocussed images). Rajagoplan and Chaudhuri [32] approach the problem by means of a space–frequency representation technique based on the Wigner distribution and the complex spectrogram.
While depth–from–defocus approximates the physical imaging process, *depth–from–focus* applies a local sharpness criterion [30]. Such a sharpness criterion could be the local variance of the intensities as proposed by the *variance method (VM)* in [38] or the *sum of modified Laplacian* as suggested in [20]. Locally, the slice with maximal sharpness is assumed to be in focus and to match the depth. Due to the absence of a deconvolution, depth–from–focus relies on the sharp information for estimating a depth value.
Regarding the recovered sharp pinhole image as a fusion of several defocused images, also *image fusion* methods are related in some sense. A simultaneous multifocus image fusion and denoising approach is proposed by Ludusan and Lavialle in [21].
Embedding the image formation model into a suitable energy functional, and posing depth–from–defocus as a variational minimisation problem is most related to our work [13, 19, 1]. While the first two approaches suggest Csiszár's information divergence as fidelity term, Aguet et al. penalise deviations from the model assumption in a quadratic way. Instead of assuming a locally equifocal surface [13] which implies a shift–invariant PSF, Jin et al. [19] refrain from such a restriction and embed a shift–variant PSF in their model.
Interpreting a shift–variant PSF as a 4-D function and a shift–invariant as a 2-D one, the approach of Aguet et al. [1], can be seen as a compromise between both. They propose the use of a 3-D PSF consisting of 2-D normalised Gaussians with varying standard deviation. On the one hand this reduces the complexity by incorporating knowledge on the depth dependence of the PSF. On the other hand, it may result in a convolution operation with

a non–normalised kernel. As a consequence, an essential physical property, namely the maximum–minimum principle w.r.t. the image intensities may be violated. This leads to a wrong model assumption especially at strong depth changes.

Our Contributions. In our present paper, we address some of the key problems of existing depth–from–defocus methods by incorporating important physical properties. In particular, the reconstruction of strong depth changes constitutes a challenge for most methods. As a remedy, we introduce a novel forward operator that closely resembles the thin lens model and preserves a maximum–minimum principle w.r.t. the unknown image intensities. We embed this operator in a variational model and minimise it with a *multiplicative* variant of the Euler–Lagrange formalism. This guarantees that the solution remains in the physically plausible positive range. Moreover, it allows a stable gradient descent evolution without the need to adapt the relaxation parameter. Synthetic and real-world experiments illustrate the advantages of our model.

This article is based on our recent conference publication [29]. However, we extend [29] in a number of important aspects: (i) We discuss the image formation models in more detail and show the relation to a standard 3–D convolution in the forward operation. For the backward operation, we demonstrate that a standard 3–D deconvolution disregards important prior knowledge and yields inferior results. Therefore, directly applying 3–D deconvolution is not a suitable approach for depth–from–defocus. (ii) While our method in [29] constitutes a proof of concept restricted to greyscale images, here we also show how to cope with multi–channel focal stacks. As a result of incorporating the information of all channels, the reconstruction quality is further improved. (iii) We analyse the impact of a robustification of the data term such that deviations of the model assumptions are penalised less severely. (iv) We extend our model to obtain a joint depth–from–defocus and denoising approach. (v) Finally, we discuss the positivity constrained minimisation strategy presented in [29] in more detail and show how it can be extended including the aspects above.

Organisation of the Paper. The goal of Section 2 is to find a forward operator that approximates the depth–of–field effect and that well fits into a variational framework. In Section 3 we formulate our variational framework which enables us to invert this imaging model. Section 4 extends our approach to a joint denoising and depth–from–defocus model. A suitable minimisation strategy is described in Section 5. Synthetic and real–world

experiments demonstrate the advantages of our approach in Section 6. The article is concluded in Section 7.

2 Image Formation Models

Before discussing the depth–from–defocus problem, we first consider the forward operation. Given the depth information and a completely sharp image, we want to generate an image as it would be produced by a camera with a limited depth–of–field. This gives us a better understanding of the depth–of–field effect and allows us to obtain the relation between depth and out–of–focus blur. For this purpose, we briefly investigate the connection between the pinhole camera and the thin lens camera model.

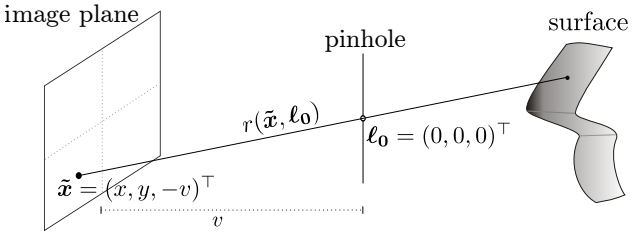

Figure 2: Pinhole camera model.

2.1 The Pinhole Camera Model

Figure 2 illustrates the *pinhole camera model*. It is the standard imaging model. It only explains geometrical optics and considers light rays as linear subsets of $\mathbb{R}^3$. For the sake of notational convenience we parametrise a ray by two points $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^3$ it passes through. Hence, a ray is fully described by

$$r(\boldsymbol{a}, \boldsymbol{b}) := \{\boldsymbol{y} \in \mathbb{R}^3 \mid \boldsymbol{y} = (1-\lambda)\ \boldsymbol{a} + \lambda\ \boldsymbol{b},\ \lambda \in \mathbb{R}\}. \tag{1}$$

We call the set of all rays $\mathcal{R}$. In the pinhole camera model, the so-called pinhole is placed in the *optical centre* $\boldsymbol{\ell}_0 := (0,0,0)^\top$ which has a distance $v \in \mathbb{R}$ to the image plane. If $\boldsymbol{x} := (x,y)^\top$ denotes the location within the image domain $\Omega_2 \subset \mathbb{R}^2$, then for each point $\tilde{\boldsymbol{x}} := (x,y,-v)^\top$ on the image plane, there exists exactly one optical ray $r(\tilde{\boldsymbol{x}}, \boldsymbol{\ell}_0) \in \mathcal{R}$ going through the pinhole. Let $d : \Omega_2 \to \mathbb{R}_+$ denote the depth map of a surface $\mathcal{S} \subset \mathbb{R}^3$, which we assume to be opaque. Then the resulting *pinhole operator* $\mathcal{F}_\mathrm{P}$ can be expressed as

$$\mathcal{F}_\mathrm{P}[\phi, d](\boldsymbol{x}) := \phi\big(Z_d\big(\tilde{\boldsymbol{x}}, \boldsymbol{\ell_0}\big)\big)\ , \tag{2}$$

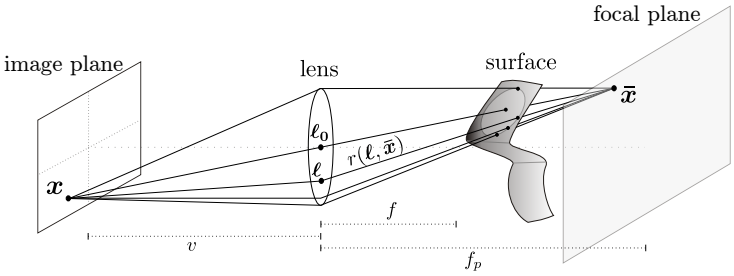

Figure 3: Thin lens camera model.

where, $\phi : \mathcal{S} \to \mathbb{R}_+$ denotes the intensity value of a surface point and $Z_d(\tilde{\boldsymbol{x}}, \boldsymbol{\ell_0})$ yields the first (i.e. closest) intersection point of the ray $r(\tilde{\boldsymbol{x}}, \boldsymbol{\ell_0})$ with the surface $\mathcal{S}$. Due to the fact that Z returns the first intersection point, it can be evaluated without knowing the complete Surface $\mathcal{S}$ but only the depth d. Since there exists at most one optical ray per image point hitting the surface, the object will be imaged completely sharp. Thus, no depth–of–field arises in this image formation model.

2.2 The Thin Lens Model

The thin lens camera model is the simplest physical model that simulates the depth–of–field effect. Instead of a pinhole, a thin circular lens with focal length f is placed in the optical center $\boldsymbol{\ell}_0$ (see Figure 3). Lens and image plane are parallel having a distance v to each other. Following the thin lens equation of geometric optics

$$\frac{1}{f_p} = \frac{1}{f} - \frac{1}{v}\,, \tag{3}$$

which is based on the intercept theorem, we obtain the distance f_p of the intersection point of the parallel ray, the central ray, and the focal ray (pinhole ray). Hence only points lying at this distance, in the so–called *focal plane* are focused sharply to the image plane [6]. Intersecting the pinhole ray $r(\tilde{\boldsymbol{x}}, \boldsymbol{\ell}_0)$ with the focal plane yields for each image point $\boldsymbol{x}$ its corresponding point within the focal plane $\bar{\boldsymbol{x}}$. This mapping corresponds to the one of the pinhole camera model.

Points not lying in the focal plane spread their intensity to a *circle of confusion* onto the image plane. In other words, if the object is not lying in the focal plane, the intensity of several surface points may blend into one single image point causing blurred information.

For an image point $\boldsymbol{x}$ the corresponding surface points are lying within the

intersection area of the bundle of lens rays with the surface. This bundle can be described using $\bar{\boldsymbol{x}}$ and all points on the lens. Following [31], one can formulate the *thin lens operator* as

$$\mathcal{F}_{\mathrm{L}}[\phi, d](\boldsymbol{x}) := \frac{1}{|\mathcal{A}|} \int_{\mathcal{A}} \phi\big(Z_d(\boldsymbol{\ell}, \bar{\boldsymbol{x}})\big) \, \mathrm{d}\boldsymbol{\ell} \, , \tag{4}$$

where $\mathcal{A}$ describes the set of points lying on the lens and $|\mathcal{A}|$ its area. The abovementioned image formation models comply with geometric optics and can be simulated using raytracing techniques [11]. However, since a large amount of blur requires processing a huge number of rays per pixel, raytracing is computationally very expensive. Therefore, researchers have been interested in finding approximations as alternatives for the simulation of photorealistic depth–of–field effects [3, 7, 33]. However, we have the requirement that the forward operation has to fit into a variational framework. So the goal of the following sections is the development of an image formation model that approximates the thin lens camera model, but can additionally be embedded into a variational framework. As a first step, we will express the thin lens camera model with the help of a spatially variant point spread function.

2.3 Spatially Variant Point Spread Function

In the pinhole camera model, each surface point is represented by exactly one image point. Instead of integrating over the lens, we can thus integrate over the sharp pinhole image u to express the thin lens camera model. With the help of a spatially variant point spread function (PSF) $H_d : \Omega_2 \times \Omega_2 \to \mathbb{R}_{0+}$ that depends on the depth profile d, the imaging process can be expressed as

$$\mathcal{F}_{\mathrm{H}}[u, d](\boldsymbol{x}) := \int_{\Omega_2} H_d(\boldsymbol{x}, \boldsymbol{y}) \; u(\boldsymbol{y}) \; \mathrm{d}\boldsymbol{y} \, , \tag{5}$$

where $\boldsymbol{x}$ describes the location within the 2-D image plane. The thin lens camera model fulfils a maximum–minimum principle w.r.t. ϕ. This guarantees that the intensity value of an image point lies between the minimum and maximum intensity value of any surface point.

$$\phi_{\min} \leq \mathcal{F}_{\mathrm{L}}[\phi, d](\boldsymbol{x}) \leq \phi_{\max} \, , \quad \forall \boldsymbol{x} \in \Omega_2 \, . \tag{6}$$

Accordingly, H_d has to preserve this w.r.t. the intensity values of the sharp pinhole image u. Therefore, we have to guarantee that for each image point $\boldsymbol{x}$ the PSF is normalised:

$$\int_{\Omega_2} H_d(\boldsymbol{x}, \boldsymbol{y}) \; \mathrm{d}\boldsymbol{y} = 1 \, , \quad \forall \boldsymbol{x} \in \Omega_2 \, . \tag{7}$$

Equation (5) can be understood as a weighted average of the sharp image intensities. Consequently, the main issue is the computation of the weights of the PSF H_d. A straightforward solution would be the use of raytracing techniques to apply the thin lens model. However, as already mentioned, raytracing is too expensive and not suitable for our variational framework. Thus, we have to find a more efficient way to approximate the weights of the PSF H_d.

2.4 Approximation of the PSF

The PSF describes the intensity distribution of each surface point onto the image plane. The intensity emitted by a surface point that does not lie within the focal plane is (uniformly) spread to a circle of confusion on the image plane [35]. The size of this circle of confusion depends on the distance d of the surface point. For simplification, let us for the moment assume that the surface is equifocal. This means that d is constant and the surface is aligned parallel to the lens. Then the circle of confusion will not change for any surface point. Moreover, Eq. (5) can be expressed in terms of a convolution. This means that one uses a spatially invariant kernel $h_d : \Omega_2 \subset \mathbb{R}^2 \to \mathbb{R}_{0+}$ instead of a spatially variant one H_d. If we additionally assume a circular lens, the PSF becomes a 2-D pillbox kernel with a radius related to the constant depth.
However, in the general case of non–constant depth map, the radius changes with the depth of each surface point. Thus, to estimate the intensity of an image point $\boldsymbol{x}$, Aguet et al. [1] weight each neighbouring point $u(\boldsymbol{y})$ corresponding to its circle of confusion, where a point having a large circle of confusion will get a small weight and vice versa. To achieve this, they introduce a 3-D PSF $h : \Omega_3 \subset \mathbb{R}^3 \to \mathbb{R}_{0+}$ as an approximation of H_d:

$$\mathcal{F}_{\mathcal{U}}[u, d](\boldsymbol{x}, z) := \int_{\Omega_2} \underbrace{h(\boldsymbol{x} - \boldsymbol{y}, z - d(\boldsymbol{y}))}_{\approx H_d(\boldsymbol{x}, \boldsymbol{y})} \, u(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y} \, , \tag{8}$$

where z represents a given focal plane. In this model, the weight not only depends on the distance of two points like convolution but also on the actual depth value. To take into account the wave character of light, Aguet et al. [1] choose a Gaussian PSF instead of a pillbox as already proposed in [28]. Then the standard deviation of the Gaussian replaces the radius of the pillbox. The approach of [1] can be illustrated as in Figure 4: First, the 2-D function h_d is lifted to a 3-D one h, composed of 2-D normalised Gaussians. The standard deviation of each Gaussian increases with increasing distance to the focal plane. Next, one cuts a slice out of this 3-D PSF h corresponding

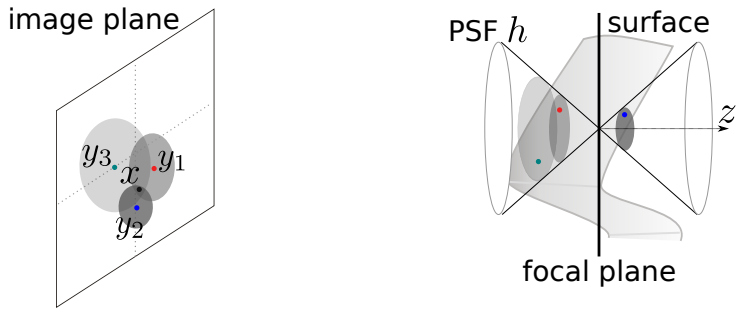

Figure 4: **Left:** Circles of confusion for different surface points appearing on the image plane. **Right:** 3-D PSF composed of 2-D normalised Gaussians.

to the depth to locally approximate H_d for a specific image point $\boldsymbol{x}$. With the help of this 3-D PSF also a second interpretation of the equation above is possible: For this, we assume that the sharp pinhole image lies in a dark volume $g : \Omega_3 \to \mathbb{R}$ corresponding to the depth profile. It can be defined as

$$g(\boldsymbol{x}, z) := u(\boldsymbol{x}) \cdot \delta(z - d(\boldsymbol{x})) \quad \text{with} \quad \delta(x) := \begin{cases} 1 & \text{if } x = 0 \,, \\ 0 & \text{else} \,. \end{cases} \tag{9}$$

Then Eq. (8) is just a standard 3-D convolution of g with the PSF h:

$$\begin{aligned} (g * h)(\boldsymbol{x}, z) &:= \int_{\Omega_3} g(\boldsymbol{y}, z') \cdot h(\boldsymbol{x} - \boldsymbol{y}, z - z') \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}z' \\ &= \int_{\Omega_3} u(\boldsymbol{y}) \cdot \delta(z' - d(\boldsymbol{y})) \cdot h(\boldsymbol{x} - \boldsymbol{y}, z - z') \, \mathrm{d}\boldsymbol{y} \, \mathrm{d}z' \\ &= \int_{\Omega_2} u(\boldsymbol{y}) \cdot h(\boldsymbol{x} - \boldsymbol{y}, z - d(\boldsymbol{y})) \, \mathrm{d}\boldsymbol{y} \,. \end{aligned} \tag{10}$$

This means that the forward operator in [1] performs a spatially invariant 3–D convolution.

2.5 Our Modification

In the equifocal case, the slice that is taken out of the PSF h is just a 2–D normalised Gaussian of a certain standard deviation. Accordingly, the maximum–minimum principle is automatically guaranteed. However, the formulation above becomes problematic if partial occlusions occur, which is expected to happen due to depth changes. In this case, the slice that is

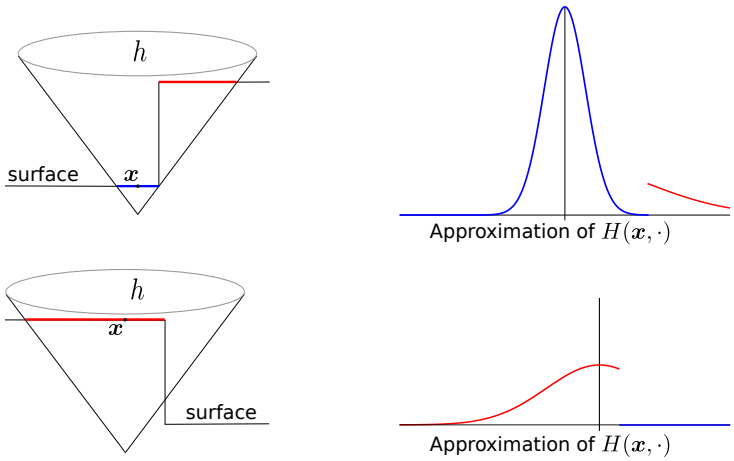

Figure 5: Unnormalised kernel. In the presence of strong depth changes, the local approximation of H_d by [1] is composed of different Gaussians (blue and red). **(a) Top row:** Overshoot: While the blue part is nearly a complete Gaussian, with the second (red) part an integration weight of 1 is exceeded. **(b) Bottom row:** Undershoot: The local composed PSF consists only of a part of a normalised Gaussian (red) and a second (blue) Gaussian already reaching negligible values. The resulting integration weight becomes smaller than 1.

cut out of h is the composition of several Gaussians with different standard deviations. As a direct consequence the normalisation cannot be guaranteed anymore. The forward operator then effectively performs spatially variant 2-D integration with an unnormalised kernel as a local approximation of H_d (see Figure 5). This results in a violation of the maximum–minimum principle w.r.t. to the image intensities.
To avoid this, we consider the local approximation of H_d and use it as a normalisation function. This introduces a novel forward operator:

$$\mathcal{F}_{\mathcal{N}}[u,d](\boldsymbol{x},z) := \frac{\mathcal{F}_{\mathcal{U}}[u,d](\boldsymbol{x},z)}{\int_{\Omega_2} h(\boldsymbol{x}-\boldsymbol{x'},z-d(\boldsymbol{x'}))\ \mathrm{d}\boldsymbol{x'}}\ . \tag{11}$$

While this normalisation may look like a small modification at first glance, it can have a large impact on the quality of the simulation of the depth–of–field effect. To demonstrate this, we compare different forward operators in Figure 6. As we can see, the result of the forward operator of [1] is very close to simple 3–D convolution (cf. Eq. (10)). Differences are caused due to the discretisation required when embedding the pinhole image into the discrete 3–D volume according to Eq. (9). Since both perform convolutions with an unnormalised kernel h on strong depth changes, they produce bright overshoots followed by dark shadows (Fig. 6(b) and (c)). This behaviour exactly matches the illustration in Figure 5. Indeed, their local violations of the maximum–minimum principle on strong depth changes produce results that are not photorealistic. This leads to a wrong model assumption. In contrast, comparing Figure 6(d) and (e) demonstrates that our normalised approach comes very close to the physically well–founded thin lens camera model which allows to create realistic depth–of–field effects.

3 Variational Formulation

In the last section, we have proposed a novel forward operator that approximates the thin lens camera model simulating the depth–of–field effect. Now we show how this operator can be inverted within a variational framework. Given a stack of blurred images, we want to jointly estimate the depth map and the completely sharp image.

3.1 Variational Model

Let us assume that a set of 2-D images capturing the same static scene with varying focal settings is arranged in a 3-D stack $s : \Omega_3 \to \mathbb{R}_+$. Here,

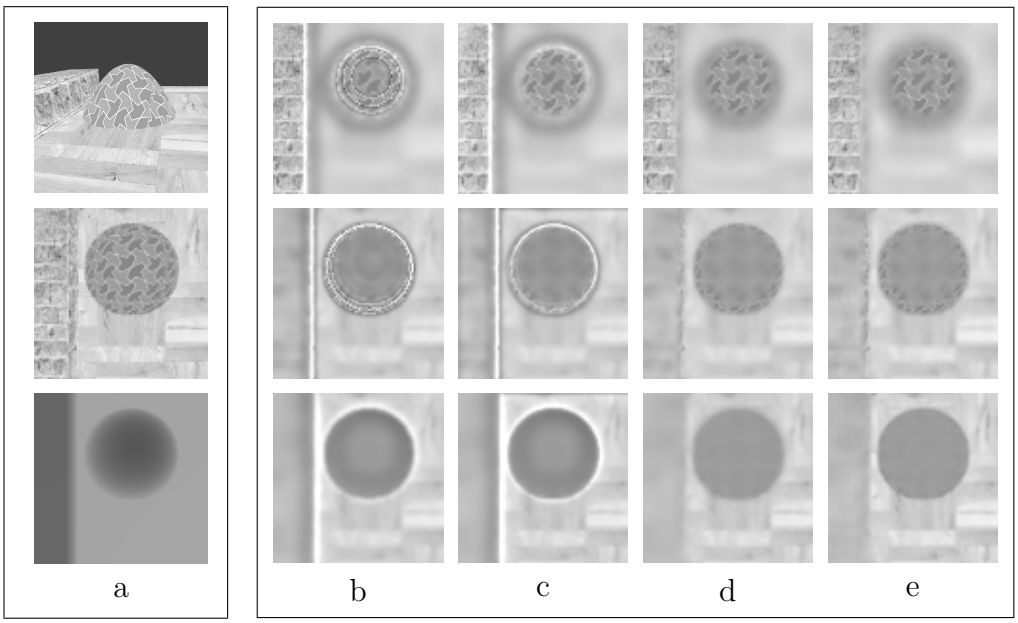

Figure 6: **(a) Top:** Synthetic 3-D test model. **(a) Middle:** Sharp image obtained with a pinhole camera renderer. **(a) Bottom:** Corresponding grey-valued coded depth map. **Right box:** Comparison of different forward operators. *From top to bottom:* Position of the focal plane changes. **(b)** Standard spatially invariant 3-D convolution (10). **(c)** Forward operator of [1] without normalisation preservation. **(d)** Our normalised forward operator (11). **(e)** Thin lens camera model (4) realised by raytracing technique.

$\Omega_3 = \Omega_2 \times \Xi$ denotes the stack volume and $\Xi \subset \mathbb{R}$ describes the set of focal settings. The sought depth map $d : \Omega_2 \to \mathbb{R}_+$ in combination with the sharp image $u : \Omega_2 \to \mathbb{R}_+$ as it would be recorded by a pinhole camera can be estimated as a minimiser of the energy

$$E(u, d) = M(u, d) + \alpha\, S(|\boldsymbol{\nabla} d|)\ . \tag{12}$$

This energy consists of a data term M and a regularisation term S with a regularisation parameter $\alpha > 0$. The data term demands similarity between the recorded focal stack and an appropriate forward operation applied to the unknown sharp image u and depth map d. Penalising the residual errors

$$r[u, d](\boldsymbol{x}, z) = s(\boldsymbol{x}, z) - \mathcal{F}_{\mathcal{N}}[u, d](\boldsymbol{x}, z) \tag{13}$$

in a quadratic way is the most common choice and results in the data term

$$M_1(u, d) = \int_{\Omega_3} \big(r[u, d]\big)^2 \mathrm{d}\boldsymbol{x}\ \mathrm{d}z\ . \tag{14}$$

This is especially suited if Gaussian distributed noise is involved. Exactly like $\mathcal{F}_{\mathcal{U}}$, our forward operator $\mathcal{F}_{\mathcal{N}}$ from (11) is linear in u but nonlinear in d, and the data term is convex in u but nonconvex in d. A minimiser of the data term alone is not unique: Regarding a homogeneous region where u is constant, it is not possible to infer an amount of blurring. Accordingly, in such regions, the depth d cannot be determined. To avoid such ambiguities and to cope with the problem of ill–posedness the regularisation term S imposes (piecewise) smoothness on the depth. This is done by penalising large gradient magnitudes of the depth profile:

$$S(|\boldsymbol{\nabla} d|) = \int_{\Omega_2} \Psi(|\boldsymbol{\nabla} d|^2) \, \mathrm{d}\boldsymbol{x} \;, \tag{15}$$

where $\boldsymbol{\nabla} := (\partial_x, \partial_y)^\top$ denotes the 2-D gradient operator and $\Psi : \mathbb{R} \to \mathbb{R}_+$ is a positive increasing function. Our experiments in Section 6 are obtained with the Whittaker–Tikhonov [39, 44] $\Psi(x^2) = x^2$ or regularised *total variation* (TV) [34] $\Psi(x^2) = \sqrt{x^2 + \epsilon}$ penaliser. We choose $\epsilon = 0.01$ to avoid singularities at $x = 0$.

3.2 Multi–Channel Images

We assume that the refraction index of the lens is independent of the wavelength of the light. Consequently, the lens treats all channels equally which results in a uniform, channel–invariant PSF. Additionally, the focal length and therewith the distance of the focal plane does not change between different channels. To approximate the depth–of–field effect given a depth map and a sharp pinhole image with channel index set $\mathcal{C}$, we thus can straightforwardly apply our forward operator $\mathcal{F}_{\mathcal{N}}$ channel–wise (cf. Figure 7).
To solve the inverse problem, Aguet et al. [1] propose to convert the multi–channel image into a single–channel one before performing their framework. However, this entails a loss of information. Instead, we believe that one can improve the reconstruction quality by incorporating the information of all channels of the recorded focal stack $\boldsymbol{s} = (s_c)_{c \in \mathcal{C}}$. Accordingly, the residual error in each channel reads:

$$r_c[\boldsymbol{u}, d] = s_c - \mathcal{F}_{\mathcal{N}}[u_c, d] \;, \quad \forall c \in \mathcal{C} \;. \tag{16}$$

To determine the sharp multi–channel image $\boldsymbol{u} = (u_c)_{c \in \mathcal{C}}$ along with the depth map d, we perform a joint penalisation of deviations by summing up over all channels $c \in \mathcal{C}$ within the data term

$$M_2(\boldsymbol{u}, d) = \int_{\Omega_3} R^2 \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}z \;, \tag{17}$$

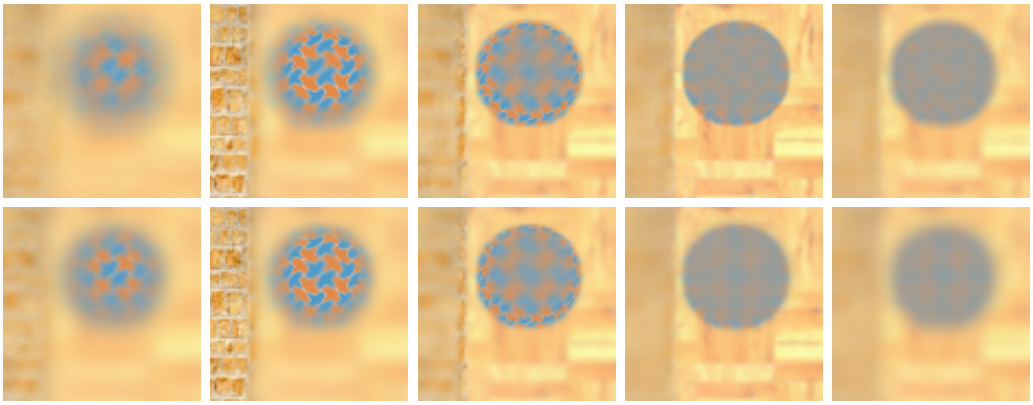

Figure 7: Multi–channel focal stack. Simulating the depth–of–field effect by applying our novel forward operator channel–wise to a RGB pinhole image. 5 out of 20 slices. **(a) Top:** Result of a thin lens camera renderer. **(b) Bottom:** Our result.

where we use the abbreviation

$$R^2 := \sum_{c \in \mathcal{C}} \big(r_c[\boldsymbol{u}, d]\big)^2 \,. \tag{18}$$

Since we estimate a joint depth map for all channels, it remains a one channel signal, and there is no change required in the smoothness term.

3.3 Robustification

Let us now have a closer look at the data term. It measures the distance between the forward operation that simulates the imaging process and the given data. Here, quadratic penalisation (least–squares) of deviations is a common choice and especially suited in the presence of Gaussian distributed noise. However, the response of an optical system to a point light source depends on a lot of different factors such as optical imperfections of the lenses or diffraction phenomena. Therefore, choosing a pillbox or Gaussian kernel can only be an approximation of the true PSF. For this reason, a quadratic penalisation of deviations may be too severe. For this reason, we follow the approach of Welk et al. [43] in the context of deconvolution and Huber [18] in the context of robust statistics. We replace the quadratic data term above by

$$M_3(\boldsymbol{u}, d) = \int_{\Omega_3} \Phi(R^2) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}z \,, \tag{19}$$

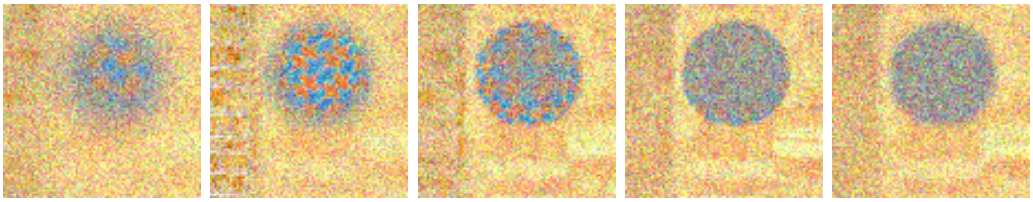

Figure 8: Focal stack disturbed by artificial Gaussian noise with standard deviation $\sigma_{\text{noise}} = 30$ and mean 0.

where we introduce a penaliser function $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ that is non–negative and subquadratic. Thus, it gives less influence to large outliers. More precisely, we apply the regularised L^1-norm $\Phi(x^2) = \sqrt{x^2 + \epsilon}$ with some small stabilisation $\epsilon = 0.01$.

4 Joint Denoising and Depth–from–Defocus

In the last section, we have proposed a variational framework for the joint reconstruction of the sharp pinhole image along with the depth map. Until now, regularisation is restricted solely to the depth map to handle the problem of ill–posedness. Since we want to recover a sharp image u, at first glance there is no need to postulate any smoothness assumption on it. However, especially in the context of microscopy at low light intensity or due to signal processing in general, the recorded stacks can contain noise. Slices of such a degraded focal stack are shown in Figure 8. In this case it can be beneficial to demand (piecewise) smoothness of the reconstructed sharp image. Of course considering denoising and depth–from–defocus as two separate tasks offers a straightforward strategy: With the help of variational image restoration [34], one is able to denoise each 2–D image $\tilde{w} : \Omega_2 \to \mathbb{R}_+$ of a focal stack before performing a standard depth–from–defocus method. In the multi–channel case where $\tilde{\boldsymbol{w}} = (\tilde{w}_c)_{c \in \mathcal{C}}$, the reconstructed 2–D image $\boldsymbol{w}$ can be determined as the minimiser of

$$E(\boldsymbol{w}) = \int_{\Omega_2} \sum_{c \in \mathcal{C}} \big(\tilde{w}_c - w_c\big)^2 \, \mathrm{d}\boldsymbol{x} + \gamma \, S_2(\boldsymbol{w}) \; . \tag{20}$$

Here the smoothness term S_2 is applied to the evolving image $\boldsymbol{w}$. In the context of deblurring noisy data, in [27, 23] a simultaneous approach is proposed. The corresponding energy can be formulated as

$$E(\boldsymbol{w}) = \int_{\Omega_2} \sum_{c \in \mathcal{C}} \big(\tilde{w}_c - k * w_c\big)^2 \, \mathrm{d}\boldsymbol{x} + \gamma \, S_2(\boldsymbol{w}) \; , \tag{21}$$

where k denotes a convolution kernel. Besides a denoising effect, regularisation declines oscillation artefacts which are typical for deconvolution methods. Since our depth–from–defocus approach also contains a deconvolution part, its obvious to follow this idea. Moreover, in our approach, we treat the unknown depth as a parameter of the blur kernel. Therefore also blind deconvolution approaches such as the one by Chan and Wong [8] can be seen as related. An extension to a joint restoration and blind deconvolution model was presented in [46, 45]. Following these approaches, we extend our method by adding a second regularisation term S_2 that is applied to the evolving sharp pinhole image $\boldsymbol{u}$:

$$E(\boldsymbol{u}, d) = M(\boldsymbol{u}, d) + \alpha\, S_1(d) + \beta\, S_2(\boldsymbol{u})\ . \tag{22}$$

Here $\alpha, \beta \geq 0$ balance the two regularisation terms

$$S_1(d) = \int_{\Omega_2} \Psi_d(|\boldsymbol{\nabla} d|^2)\ \mathrm{d}\boldsymbol{x}\ , \tag{23}$$

$$S_2(\boldsymbol{u}) = \int_{\Omega_2} \Psi_u(|\boldsymbol{\nabla} \boldsymbol{u}|^2)\ \mathrm{d}\boldsymbol{x} \tag{24}$$

against the data term, where $|\boldsymbol{\nabla} \boldsymbol{u}|^2 := \sum_{c \in \mathcal{C}} |\boldsymbol{\nabla} u_c|^2$, and the functions Ψ_d, Ψ_u allow different smoothing behaviours for each task.

5 Minimisation

Afterwards we have presented our novel variational depth–from–defocus functional, let us now investigate the search for a suitable minimiser. For this purpose, one commonly follows classical additive *Euler-Lagrange* formalism. In this section, however, we want to give also a brief insight into the less known multiplicative Euler–Lagrange variant and show its advantages. We complete this section by giving numerical and implementation details. For the sake of notational simplicity, we restrict ourself to grey value data sets. The extension to colour data sets will be straightforward.

5.1 Euler-Lagrange Equations

For a one–channel focal stack s our joint variational model reads

$$E(u, d) = \int_{\Omega_3} \Phi\Big(\big(\underbrace{s - \mathcal{F}_{\mathcal{N}}[u, d]}_{=: r[u,d]} \big)^2 \Big)\ \mathrm{d}\boldsymbol{x}\ \mathrm{d}z + \int_{\Omega_2} \Big(\alpha\ \Psi_d(|\boldsymbol{\nabla} d|^2) + \beta\ \Psi_u(|\boldsymbol{\nabla} u|^2) \Big) \mathrm{d}\boldsymbol{x}\ . \tag{25}$$

From variational calculus one knows that a sharp pinhole image u and depth map d as a joint minimiser of (25) must necessarily fulfil the Euler–Lagrange equations

$$\frac{\delta E}{\delta u} = 0 \quad \text{and} \quad \frac{\delta E}{\delta d} = 0 \tag{26}$$

and its corresponding natural boundary conditions. For this purpose, one has to derive the variational gradients $\frac{\delta E}{\delta u}$ and $\frac{\delta E}{\delta d}$. To compute a variational gradient $\frac{\delta E}{\delta u}$ one commonly follows the classical (additive) Euler–Lagrange formalism [15] and requires

$$\left\langle \frac{\delta E}{\delta u}, v \right\rangle \stackrel{!}{=} \frac{\partial}{\partial \epsilon} E(u + \epsilon\, v)\Big|_{\epsilon=0}, \quad \forall v \in \mathbb{R} \to \mathbb{R} \tag{27}$$

where $\langle \cdot, \cdot \rangle$ denotes the standard inner product. The right hand side of (27) can be interpreted as a directional derivative in the direction of a function v where the perturbation acts additively. Eventually, we obtain

$$\frac{\delta E}{\delta u}(\boldsymbol{x}) = -2 \left(\big(\Phi' \cdot \overline{r}\big) * h^* \right)(\boldsymbol{x}, d(\boldsymbol{x})) - 2\beta \cdot \operatorname{div}\Big(\Psi'_u \; \boldsymbol{\nabla} u \Big) , \tag{28}$$

where we have introduced the abbreviations $\Phi' := \Phi'(R^2)$, $\Psi'_u := \Psi'_u(|\boldsymbol{\nabla} u|^2)$, and $\overline{r} := N^{-1} \cdot r$. The residuum (16) is given by r, and N corresponds to the normalisation function, i.e. the denominator in (11). The operator $*$ expresses a 3-D convolution and $h^*(x)$ denotes the adjoint of h by $h^*(x) := h(-x)$. Following the same formalism w.r.t. the depth d, we obtain the variational gradient

$$\begin{aligned} \frac{\delta E}{\delta d}(\boldsymbol{x}) = 2 \left(\big(\Phi' \cdot \overline{r}\big) * h_z^* \right)(\boldsymbol{x}, d(\boldsymbol{x})) \cdot u - 2 \left(\big(\Phi' \cdot \overline{r} \cdot \mathcal{F}_{\mathcal{N}}[u, d]\big) * h_z^* \right)(\boldsymbol{x}, d(\boldsymbol{x})) \\ - 2\alpha \cdot \operatorname{div}\Big(\Psi'_d(|\boldsymbol{\nabla} d|^2) \; \boldsymbol{\nabla} d \Big) , \end{aligned} \tag{29}$$

where h_z^* denotes the partial derivative of h^* in z–direction. Since $\Psi'_u, \Psi'_d > 0$, the natural boundary conditions read

$$\boldsymbol{n}^\top \boldsymbol{\nabla} u = 0 \qquad \text{and} \qquad \boldsymbol{n}^\top \boldsymbol{\nabla} d = 0 , \tag{30}$$

where $\boldsymbol{n}^\top$ is the normal vector at the image boundary.

5.2 Enforcing Positivity

Variational depth–from–defocus states an ill–posed problem with a non–unique minimiser. To reduce the problem of ill–posedness a common remedy

is given by imposing additional inequality constraints. In this way, the solution can be restricted to only physically plausible values. Regarding the imaging process, we know that the number of photons arriving at the image sensor is larger than zero. Hence, the intensity values of each channel has to be positive. Also the considered surface is lying in front of the lens which implies a positive range concerning the depth values. While the classical additive Euler–Lagrange formalism *does not* impose any constraints on the estimation, the *multiplicative* Euler–Lagrange formalism offers an interesting alternative to retain the positivity of the solution [42]. Based on a perturbation that acts in a multiplicative way instead of an additive one, we obtain

$$\left\langle \frac{\delta^* E}{\delta u}, v \right\rangle \stackrel{!}{=} \frac{\partial}{\partial \epsilon} E(u + \epsilon u \cdot v)\Big|_{\epsilon=0}, \quad \forall v \in \mathbb{R} \to \mathbb{R} \tag{31}$$

for the sharp pinhole image. Applying the same formalism w.r.t. d, this results in the following functional derivatives:

$$\frac{\delta^* E}{\delta u} = u \cdot \frac{\delta E}{\delta u} \quad \text{and} \quad \frac{\delta^* E}{\delta d} = d \cdot \frac{\delta E}{\delta d}. \tag{32}$$

The boundary conditions remain the same as in the additive formalism.
In [42], Welk illustrates in two different ways, why the multiplicative Euler–Lagrange formalism restricts the solution to positive values: On the one hand, it can be shown that the multiplicative Euler–Lagrange formalism corresponds to the reparametrisation $u = \exp(w)$ and $d = \exp(z)$. On the other hand, one can also observe that the multiplicative functional gradients $\frac{\delta^* E}{\delta u}$ and $\frac{\delta^* E}{\delta d}$ occur within the additive formalism when one replaces the Euclidean metric $\mathrm{d}u$ by a hyperbolic one, i.e. $\mathrm{d}u/u$. Thus, one effectively moves unwanted values to infinite distance.

5.3 Discretisation and Implementation

The benefits of applying the multiplicative Euler–Lagrange formalism are two–fold. In the context of quality, it constrains the solution to the plausible positive range. As a second benefit, the multiplicative Euler–Lagrange formalism also gives us access to an efficient semi–implicit iteration scheme. Concerning the multiplicative gradient from Equation (31) w.r.t. u, we suggest the following semi–implicit scheme:

$$\frac{u^{k+1} - u^k}{\tau} = 2\, u^{k+1}\Big(\big(\Phi'^k \cdot \overline{r}^k\big) * h^*\Big)(\boldsymbol{x}, d) + 2\beta \cdot \mathrm{div}\Big(\Psi_u'^k\, \boldsymbol{\nabla} u^{k+1}\Big) \cdot u^k\,, \tag{33}$$

where τ is the relaxation parameter and k denotes the iteration level. Furthermore, we have used the abbreviations $\Phi'^k := \Phi'(R^{2^k})$, $\Psi_u'^k := \Psi_u'(|\boldsymbol{\nabla} u^k|^2)$.

For the estimation of the depth map d, we obtain

$$\begin{aligned}\frac{d^{k+1}-d^k}{\tau} = -2 \Bigg(& \Big(\big(\Phi'^k \cdot \overline{r}^k\big) * h_z^* \Big)(\boldsymbol{x}, d^k) \cdot u \\ & - \Big(\Big(\Phi'^k \cdot \overline{r}^k \cdot \mathcal{F}_{\mathcal{N}}[u, d^k] \Big) * h_z^* \Big)(\boldsymbol{x}, d^k) \Bigg) \cdot d^{k+1} \\ & + 2\alpha \cdot \operatorname{div}\Big(\Psi_d'(|\boldsymbol{\nabla} d^k|^2) \; \boldsymbol{\nabla} d^{k+1} \Big) \cdot d^k \; . \end{aligned} \tag{34}$$

Using the standard additive Euler–Lagrange formalism requires to adapt the relaxation parameter in each iteration step. For this purpose, the computation of a suitable step–size has to be done by expensive algorithms such as the backtracking line–search method (see e.g. [1]). We found that the suggested semi–implicit scheme above gives a higher stability range w.r.t. the relaxation parameter, the step–size can be chosen fixed in advance. Therefore, we can refrain from such expensive algorithms.
To apply the presented method on a focal stack consisting of N_z digital images each sampled on a rectangular regular grid of size $N_x \times N_y =: N$, we replace continuous functions by their discrete counterparts. Hence, in each pixel $(i,j) \in \{1,...,N_x\} \times \{1,...,N_y\}$, we have to fulfil

$$\frac{u_{i,j}^{k+1} - u_{i,j}^k}{\tau} = \underbrace{2 \left[\Big(\big(\Phi'^k \cdot \overline{r}^k\big) * h^* \Big) \right]_{i,j,d_{i,j}}}_{\boldsymbol{D}_1} \cdot u_{i,j}^{k+1} + \beta \cdot \underbrace{2 \Big[\operatorname{div}\Big(\Psi_u'^k \; \boldsymbol{\nabla} u^{k+1} \Big) \Big]_{i,j}}_{\boldsymbol{A}(\boldsymbol{u}^k)\boldsymbol{u}^{k+1}} \cdot \underbrace{u_{i,j}^k}_{\boldsymbol{D}_2} \tag{35}$$

w.r.t. the sharp pinhole image u. To implement the 3–D convolution, we transfer its components into the Fourier domain and use the convolution theorem. While Equation (33) can be solved directly for $\beta = 0$, we have to solve a linear system of equations of the form $\boldsymbol{B}\boldsymbol{x} = \boldsymbol{b}$ otherwise. By introducing a single-index notation (e.g. row–major ordering) and arranging 2–D images $u : \mathbb{R}^2 \to \mathbb{R}_+$ in vectors $\boldsymbol{u} \in \mathbb{R}^N$, we can express the point–wise multiplication with the help of diagonal matrices $\boldsymbol{D}_1, \boldsymbol{D}_2 \in \mathbb{R}^{N \times N}$. Furthermore, the discrete regularisation term can be written with the help of a matrix–vector multiplication $\boldsymbol{A}(\boldsymbol{u}^k)\boldsymbol{u}^{k+1}$. Therewith the above equation turns into

$$\underbrace{\big(\boldsymbol{I} - \tau(\boldsymbol{D}_1(\boldsymbol{u})^k + \beta \cdot \boldsymbol{D}_2(\boldsymbol{u}^k) \cdot \boldsymbol{A}(\boldsymbol{u}^k))\big)}_{\boldsymbol{B}} \cdot \underbrace{\boldsymbol{u}^{k+1}}_{\boldsymbol{x}} = \underbrace{\boldsymbol{u}^k}_{\boldsymbol{b}} \; . \tag{36}$$

To solve this system of equations we choose the Jacobi algorithm:

$$\boldsymbol{x}^{m+1} = \boldsymbol{\mathcal{D}}^{-1}(\boldsymbol{\mathcal{T}}\boldsymbol{x}^m + \boldsymbol{b}), \quad \boldsymbol{x}^0 = \boldsymbol{u}^k, \quad m = 0, 1, 2, .. \tag{37}$$

where

$$\boldsymbol{\mathcal{T}} = \tau \cdot \beta \cdot \boldsymbol{A}_{\text{rest}} \boldsymbol{D}_2 \ ,$$
$$\boldsymbol{\mathcal{D}} = \boldsymbol{I} - \tau \cdot \boldsymbol{D}_1 - \tau \cdot \beta \cdot \boldsymbol{A}_{\text{diag}} \boldsymbol{D}_2 \ , \tag{38}$$

and we split $\boldsymbol{A}(\boldsymbol{u^k}) = \boldsymbol{A}_{\text{rest}} + \boldsymbol{A}_{\text{diag}}$ into an easily invertible diagonal part and the remaining off–diagonal one. Concerning the estimation of the depth map d, we proceed analogously by setting $\boldsymbol{D}_2 = d^k_{i,j}$, exchanging β by α and setting

$$\boldsymbol{D}_1 := -2\Big[\big(\Phi'^k \cdot \overline{r}^k\big) * h_z^*\Big]_{i,j,d_{i,j}} \cdot \ u_{i,j} + 2\Big[\Big(\Phi'^k \cdot \overline{r}^k \ \cdot \mathcal{F}_{\mathcal{N}}[u, d^k]\Big) * h_z^*\Big]_{i,j,d_{i,j}} \ .$$

Since both systems of equations depend on each other, we perform an alternating minimisation strategy [22]. Solving the first problem (e.g. recovering the sharp image u), the second problem (e.g. the estimation of the depth d) remains fixed. After a fixed number of gradient descent steps, the roles are exchanged. Such a strategy is very common e.g. in blind deconvolution problems [8] where both the sharp image and the blur kernel have to be estimated. To handle the problem of nonconvexity, we apply a coarse–to–fine strategy where the solution of the coarser level provides the initialisation for the next finer one.
To guarantee the positivity of our solution $\boldsymbol{x}^{k+1}$ under the condition that $\boldsymbol{x}^k$ is positive, we have to restrict the relaxation parameter τ. Since we perform isotropic regularisation, we know that $\mathcal{T} > 0$ for all $\tau > 0$. We thus only have to consider $\mathcal{D}$. To remain in a positive range

$$\mathcal{D}_{\ell,\ell} > 0 \quad \Leftrightarrow \quad [-\boldsymbol{D}_1 - \beta \cdot \boldsymbol{D}_2 \cdot \boldsymbol{A}_{\text{diag}}]_{\ell,\ell} > -\frac{1}{\tau} \tag{39}$$

has to hold $\forall \ell \in \{1, \ldots, N\}$. For the texture as well as for the estimation of the depth map $\boldsymbol{D}_1$ can have an arbitrary sign. In case that all entries of $\boldsymbol{D}_1$ are negative Eq. (39) is fulfilled for all τ, since $-\beta \cdot \boldsymbol{D}_2 \cdot \boldsymbol{A}_{\text{diag}} > 0$. Otherwise, the solution remains positive for

$$\tau < \frac{-1}{\min_\ell [-\boldsymbol{D}_1]_{\ell,\ell}} \ . \tag{40}$$

6 Experiments

To compare the performance and reconstruction quality of each of our presented concepts, in this section, we apply our approaches on synthetic and

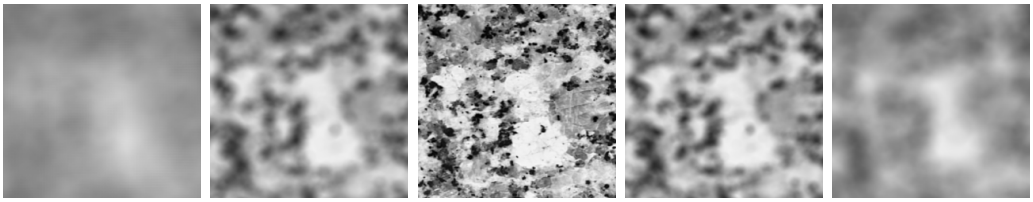

Figure 9: Equifocal case: High textured plane parallel to the lens. 5 out of 9 images of a focal stack rendered by a thin lens camera renderer.

real world data. The main focus lies on comparing different forward operators, the benefit of robustification as well as the usage of full colour information. Besides that, we want to demonstrate the capabilities of our novel joint depth-from-defocus and denoising approach.

6.1 Synthetic Data

As already mentioned in Section 2.4, the forward operator of Aguet et al. can be understood as a standard 3-D convolution when placing the sharp pinhole image in a dark volume corresponding to the depth profile. Consequently, the natural question arises whether a 3-D deconvolution can be used to recover the sharp pinhole image from a recorded focal stack s. Our first experiment deals with exactly this question. To this end, we generate a focal stack with the help of a thin lens camera renderer (lens diameter $D = 2.69$ cm, lens distance to image plane $v = 35$ mm), where the distance of the focal plane to the lens varies equidistantly from $f_p = 3$ cm to $f_p = 11$ cm. We produce 9 images of size 250×250 in total (see Figure 9). The 3-D model that we consider in this case is simply a highly textured equifocal plane at distance $d(\boldsymbol{x}) = 7$ cm. We deblur the generated focal stack with variational deconvolution using an extension of Equation (21) to the three–dimensional case where the 3-D PSF is given by h. Regarding Figure 10 (a) one recognises that variational 3-D deconvolution is not able to reconstruct the sharp slice in a reasonable way. This is because the standard 3D deconvolution does not model the fact that the focal stack has to originate from a dark volume with only a single sharp slice according to a depth profile. However, if we incorporate depth information e.g. by setting all values to zero that not correspond to the actual depth in each iteration, we obtain the result in Figure 10(b). Of course this is not a practical solution since it requires knowledge of the correct depth. Since such an approach is not useful even in this simple scenario, we do not consider the standard 3–D deconvolution any further.

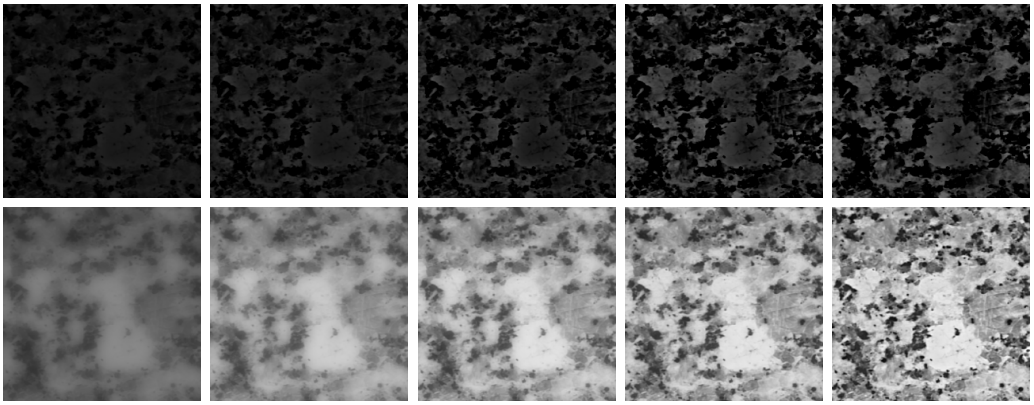

Figure 10: Variational 3-D deconvolution applied to the focal stack of Figure 9. The slice corresponding to the true depth $d(\boldsymbol{x}) = 7$ cm at different iteration time is shown. Using gradient descent methods, the iteration time denotes the product of time step size and actual iteration number. **(a) Top:** Without further assumptions. *From left to right:* Iteration time = 0.25, 0.50, 1.00, 5.00, and 10.00. **(b) Bottom:** Additional all slices not corresponding to depth profile are set to zero. *From left to right:* Iteration time = 0.10, 0.25, 0.50, 1.00, and 5.00.

In our second experiment, we compare our approach against the variance method and the approach of Aguet et al. [1]. To this end, we use the thin lens camera renderer with the same optical settings as in the first experiment. We move the focal plane equidistantly from $f_p = 3$ cm to $f_p = 7$ cm to render 20 images of the 3-D model from Fig 6(a). First, we restrict ourselves to the one–channel case, data term M_1 and Whittaker–Tikhonov regularisation on d. We refrain from any regularisation of the sharp pinhole image u. In Figure 6(e) 3 different slices of this focal stack can be seen. Figure 11 shows results of the different approaches: The variance method (Fig. 11(a)) suffers from two undesired hills in front of and behind the hemisphere. This is because the variance method misinterprets the blur circle of the hemisphere as a higher local sharpness than the flat contrast of the textured floor.
Figure 11(b) and (c) demonstrate the consequences ignoring normalisation. Applying $\mathcal{F}_U$ to a depth map produces severe over- and undershoots at strong depth changes. This implies that keeping strong depth changes in the inverse operation drastically increases the residual error at those locations. Thus, when minimising the residual error with $\mathcal{F}_U$ as forward operator, smooth changes of the depth are preferred. This can be seen as an unwanted regularisation of the depth. Furthermore, comparing Figure 11(b) and (c) one

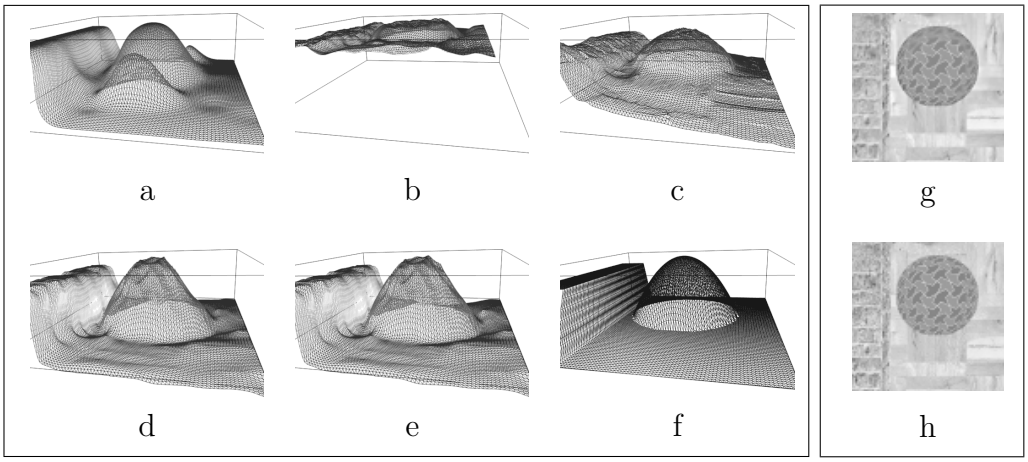

Figure 11: Comparison of different reconstruction methods. **Left box:** *In reading order* **(a)** Variance Method (VM) with a subsequent Gaussian smoothing step (patch-size = 6, $\sigma = 4.0$). **(b)** Without normalisation $\mathcal{F}_{\mathcal{U}}$, initialised with constant depth ($\alpha = 45$). **(c)** Dito, initialisation provided by the Variance Method. **(d)** Our normalised approach $\mathcal{F}_{\mathcal{N}}$, initialised with constant depth ($\alpha = 150$). **(e)** Dito initialisation provided by the Variance Method. **(f)** Ground truth of the depth profile. **Right box: (g) Top:** Estimated pinhole image belonging to the depth map (e). **(h) Bottom:** Ground truth of the pinhole image.

observes that the result is strongly affected by the initialisation. While in the first one a constant depth map is used, the second one was initialised with the result of the variance method. Due to the strong regularisation implied by $\mathcal{F}_U$, initialising with a constant depth, the method is not able to converge to a reasonable solution.
In contrast to that, our forward operator $\mathcal{F}_N$ approximates the thin lens camera model. Thus, it is much closer to the physical imaging process, especially at strong depth changes. Embedding our forward operation into a variational framework improves the estimation of the unknown depth as well as the sharp pinhole image substantially. Indeed, as we can see in the Figures 11(d) and (e), the hemisphere as well as the strong depth change at the wall are well reconstructed and no smoothing effect implied by our forward operator exists. Also regarding the sharp image (Figure 11(g)), our results match the ground truth in a better way. This can also bee seen in Table 1. Moreover, the initialisation of our approach does not severely affect the solution.

In our third Experiment, we investigate a possible improvement of our re-

Table 1: Error measurement. To judge the estimated pinhole image as well as the depth map against their ground truth, we consider the mean squared error (MSE) and the structural similarity (SSIM) [40]. We show the error of the pure variance method (VM) as well as the one with an additional post–smoothing step with variance σ. Further, the operator $\mathcal{F}_{\mathcal{U}}$ and our normalised imaging model $\mathcal{F}_{\mathcal{N}}$ is considered. The latter two are initialised once with a constant depth and once with an estimation of the VM.

Method		VM		$\mathcal{F}_{\mathcal{U}}$		Ours	
		$\sigma = 0$	$\sigma = 4$	const.	VM	const.	VM
Depth	**MSE**	2.83	1.10	21.66	2.26	0.77	**0.58**
	SSIM	0.976	0.995	0.94	0.992	0.995	**0.997**
Image	**MSE**	48.33	46.38	124.52	52.17	49.09	**45.17**
	SSIM	0.87	0.87	0.67	0.87	0.90	**0.92**

sult of Figure 11(e) by replacing Whittaker–Tikhonov regularisation [39, 44] $\Psi(x^2) = x^2$ by regularised TV [34] $\Psi(x^2) = \sqrt{x^2 + \epsilon}$ with $\epsilon = 0.01$. Regarding Figure 12, we can clearly see that depending on the smoothness weight α a better reconstruction of the wall is possible. However the cost is a cropped hemisphere suffering from a too steep depth change. This can also be seen regarding the reconstruction of the texture of the hemisphere. A quantitative comparison for different α is provided in Table 2.

The fourth experiment investigates the benefit of incorporating all channels

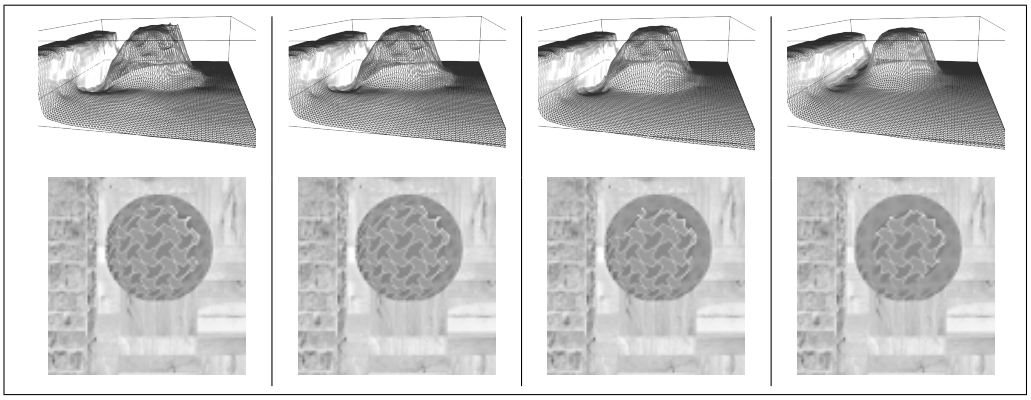

Figure 12: Visual comparison of quadratic data term with TV regularisation. *From left to right:* **(a):** $\alpha = 150$. **(b):** $\alpha = 180$. **(c):** $\alpha = 220$. **(d):** $\alpha = 300$.

Table 2: Quantitative comparison for different α using TV regularisation. Apart from TV regularisation, the same settings as in our first experiment (Figure 11(e)) are applied.

Parameter α		150	180	220	300
Depth	**MSE**	0.51	**0.48**	0.77	1.06
	SSIM	0.9959	**0.9962**	0.9944	0.9933
Image	**MSE**	53.41	**51.29**	55.80	54.82
	SSIM	0.9115	**0.9125**	0.8974	0.8818

given a multi–channel focal stack. To this end, we apply our algorithm with data term M_2 and Whittaker–Tikhonov regularisation on d to an RGB version of the focal stack from the second experiment (see Figure 7(a)). In Figure 13 its result is compared against our single channel approach. As we can recognise, incorporating the information of all channels not only leads to a more appealing and accurate estimation of the sharp pinhole image, but also the reconstruction of the depth map is closer to the ground truth.

The fifth experiment illustrates the impact of the robustification of the data term. We compare the results using a quadratic data term M_2 against the data term M_3 that penalises outliers less severely. In Figure 14 both

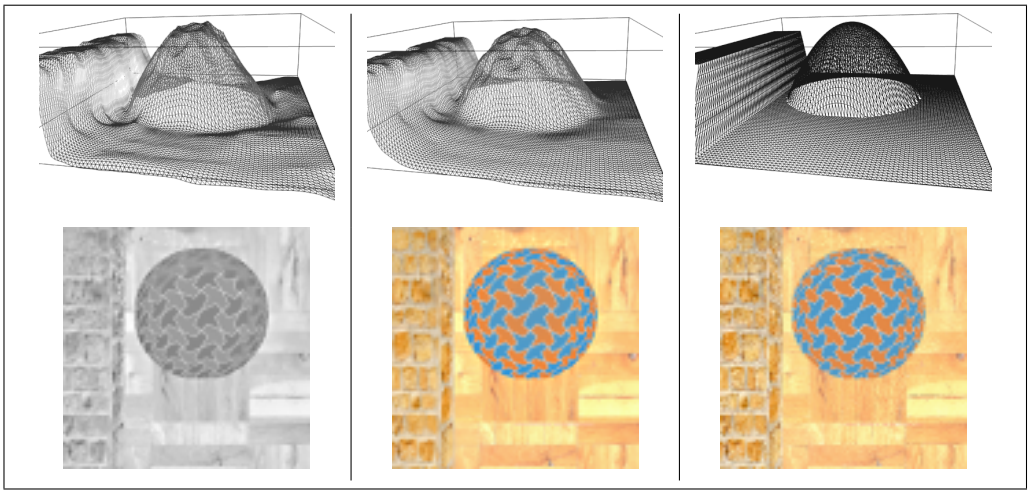

Figure 13: Visual comparison. **(a) Left:** Single channel approach. **(b) Middle:** Incorporating all channels of a RGB focal stack. **(c) Right:** Ground truth.

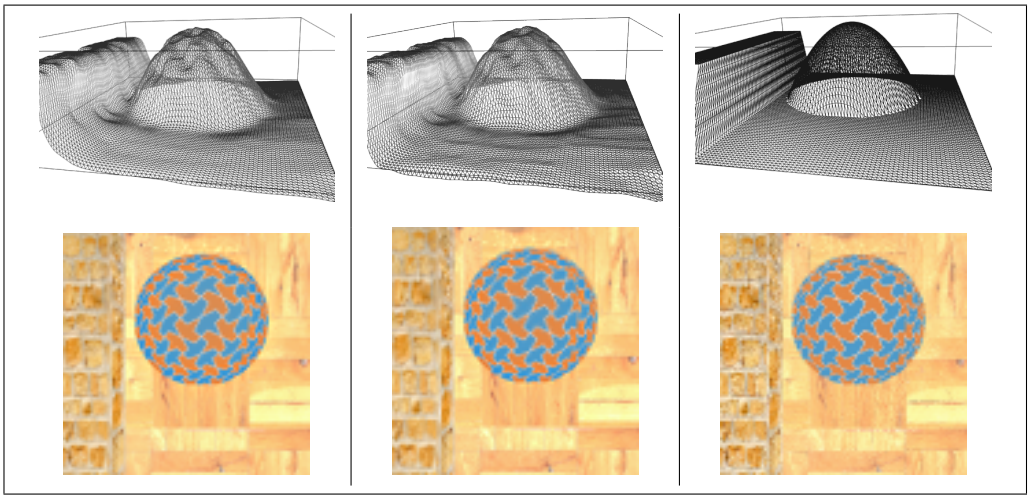

Figure 14: **(a) Left:** Without robustification. **(b) Middle:** With robustification. **(c) Right:** Ground truth.

results are shown. The data term M_3 gives a higher reconstruction quality especially at the strong depth change at the wall (cf. Figure 15). Next, we replace Whittaker–Tikhonov by TV regularisation on d. The results are shown in Figure 16. As we can see, in contrast to our second experiment, this time TV can improve the reconstruction result although the hemisphere is shrunk. Table 3 summarises the results.

In our next experiment, we investigate the reconstruction performance if only half the number of slices of the focal stack are available. To this end, we remove the odd number of slices of our colour focal stack. Next, we apply our method using data term M_3 as well as Whittaker–Tikhonov or TV regularisation. The results are compared in Figure 17 and Table 4.

The last experiment on synthetic data demonstrates the potential of our

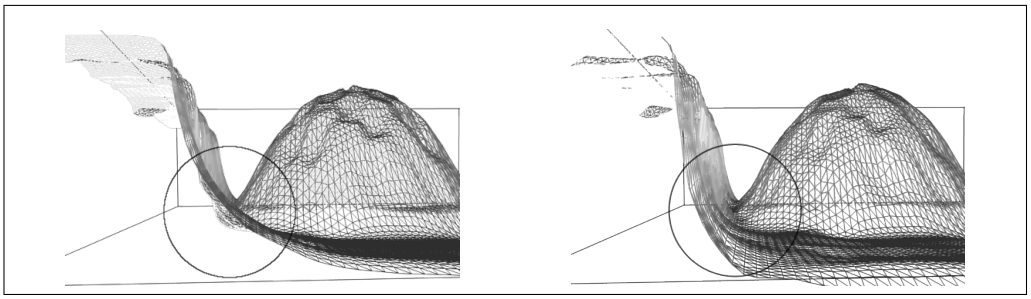

Figure 15: **(a) Left:** Without robustification. **(a) Right:** With robustification.

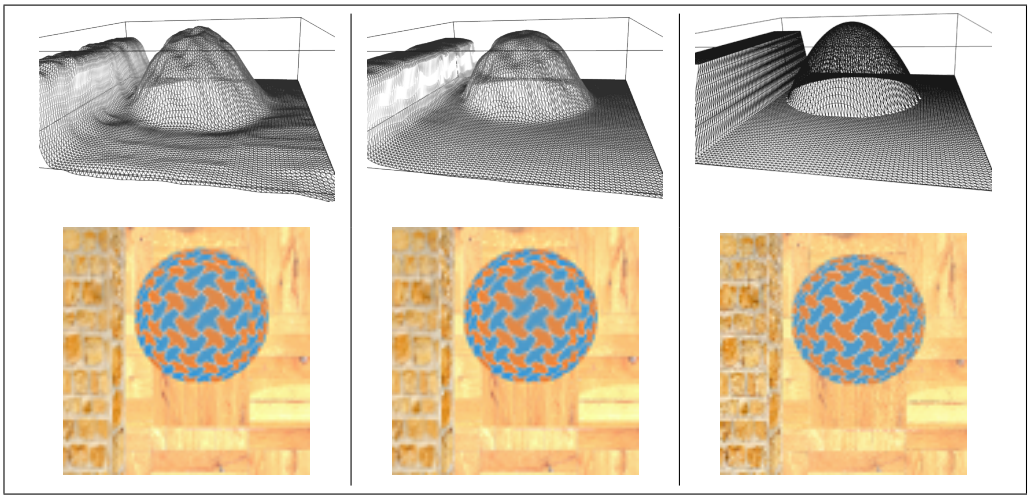

Figure 16: **(a) Left:** Robust with Whittaker–Tikhonov regularisation. **(b) Middle:** Robust with TV regularisation. **(c) Right:** Ground truth.

Table 3: Quantitative comparison: Colour, robustification and TV regularisation. Again mean squared error (MSE) and the structural similarity (SSIM) [40] are used. We consider the benefit of incorporating all image channels as well as the influence of a robustification. We use the sharp grey scale pinhole image as ground truth. Therefore, we convert the reconstructed sharp colour image to grey scale before measuring the MSE and SSIM.

Method		$\mathcal{F}_{\mathcal{N}}$	Colour ($\alpha = 780$)	Colour + Robust ($\alpha = 60$)	Colour + Robust + TV ($\alpha = 20$)
Depth	**MSE**	0.58	0.23	0.17	**0.11**
	SSIM	0.9973	0.9985	0.9987	**0.9990**
Image	**MSE**	45.17	**33.73**	40.61	40.77
	SSIM	0.9175	**0.9301**	0.9140	0.9150

novel joint denoising and depth–from–defocus approach. To realise this, we perform Charbonnier regularisation [9] $\Psi_u(x^2) := 1/\sqrt{1 + x^2/\lambda^2}$ with $\lambda = 2$ on u and Whittaker–Tikhonov regularisation on d. We consider the focal stack of the second experiment and add artificial Gaussian noise with zero mean and standard deviation $\sigma_{\text{noise}} = 30$ (cf. Figure 8). As a baseline for comparison, we consider a sequential framework where each slice of the focal stack is denoised by image restoration (corresponding Eq. (20)) in advance

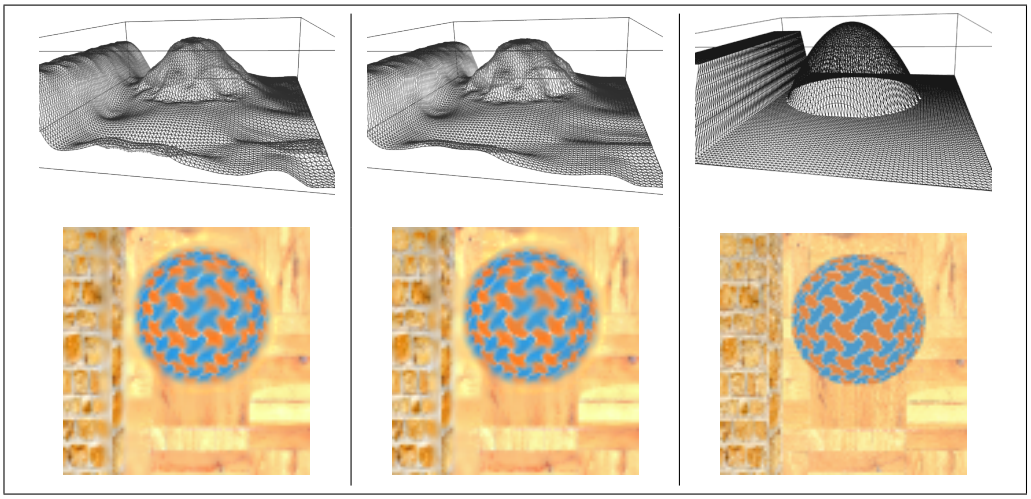

Figure 17: Reconstruction result if only half number of slices are available. **(a) Left:** With Whittaker–Tikhonov regularisation. **(b) Middle:** With TV regularisation. **(c) Right:** Ground truth.

Table 4: Quantitative comparison if only half number of slices are available.

Method		Whittaker–Tikhonov ($\alpha = 32$)	TV ($\alpha = 6.0$)
Depth	**MSE**	0.27	**0.26**
	SSIM	0.9969	**0.9971**
Image	**MSE**	**80.26**	81.72
	SSIM	0.8392	**0.8407**

before performing depth–from–defocus. We optimised the parameters α and β, γ in order to minimise the MSE w.r.t. reconstructed pinhole image u and depth map d respectively. The behaviour of the MSE of u and d in dependency of the chosen α and β, γ are shown in Figure 19 for the sequential approach and Figure 20 respectively for our novel joint approach. Since our data set is disturbed by strong Gaussian noise, we refrain from robustification in this experiment and consider a quadratic penalisation which is tailored for this kind of noise. The comparisons in Figure 18 and Table 5 show that our novel joint approach outperforms the sequential one qualitatively and quantitatively.

Table 5: Quantitative comparison of sequential and joint approach.

Method		Seq. ($\alpha = 500$, $\gamma = 20$)	Joint ($\alpha = 350$, $\beta = 2.0$)
Depth	**MSE**	0.52	**0.32**
	SSIM	0.9971	**0.9976**
Image	**MSE**	110.02	**103.18**
	SSIM	0.7421	**0.7649**

6.2 Real–World Data

Until now the experiments have been restricted to synthetic data. While synthetic experiments offer the advantage of having a ground truth and therewith the possibility to judge the results quantitatively, it is also important to consider the results on focal stacks captured by a real optical system. In Figure 21(a), we can see a focal stack showing a house fly eye. For this experiment, our approach uses a coarse–to–fine strategy. On the coarsest grid the method is initialised with the variance method. Since, there is no noise involved and we want a sharp reconstruction of the pinhole image, we refrain from regularisation on u. On d we apply TV regularisation. Figure 21(b) shows the estimated sharp pinhole image along with the estimated depth map. The determined pinhole image is reconstructed well, and the

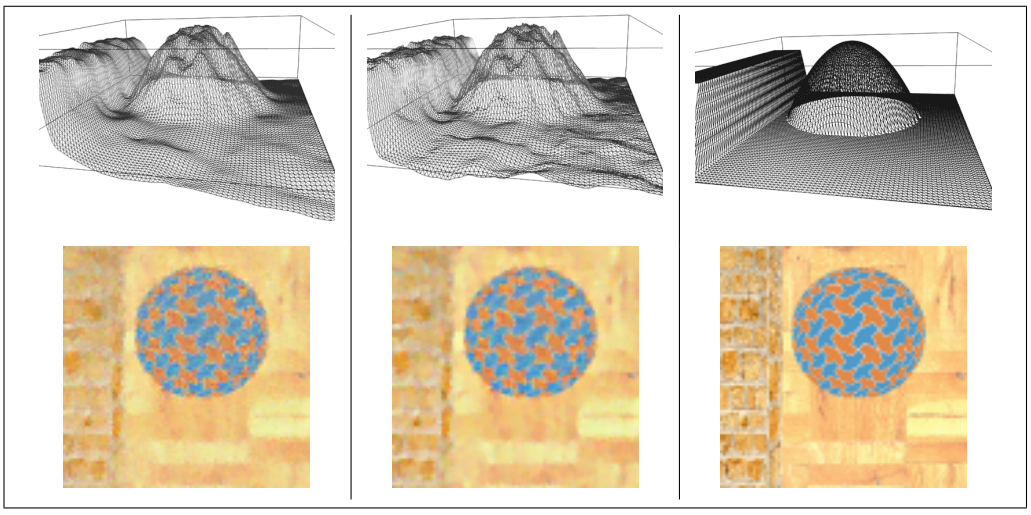

Figure 18: **(a) Left:** Sequential approach. **(b) Middle:** Joint approach. **(c) Right:** Ground truth.

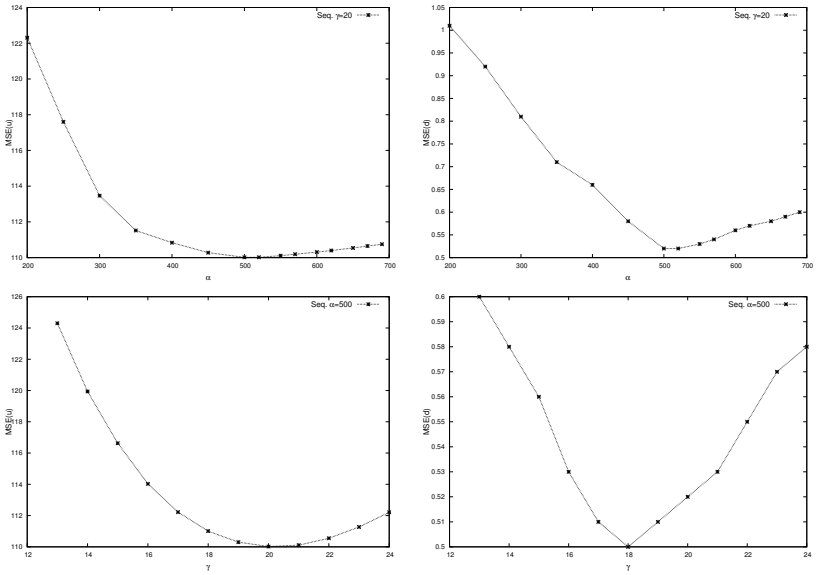

Figure 19: Sequential approach: MSE of u and d respectively w.r.t. the parameter α (first row) and γ (second row).

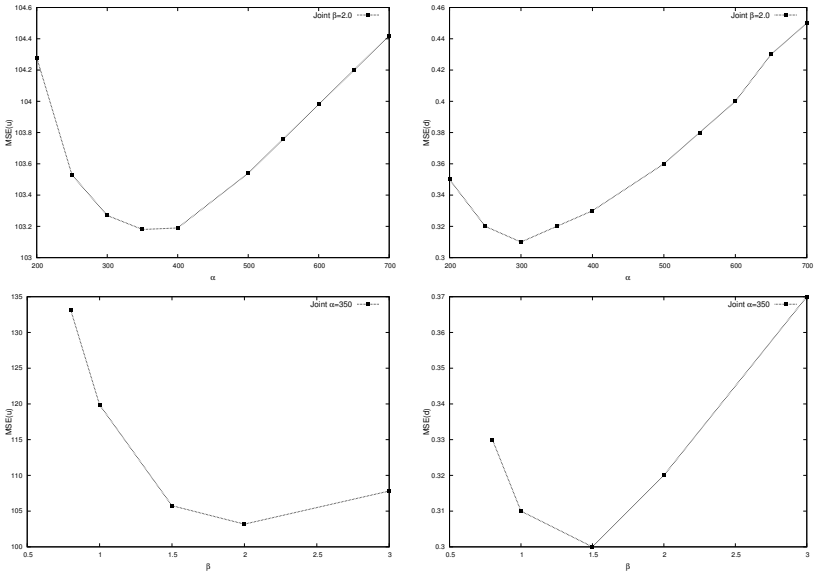

Figure 20: Joint approach: MSE of u and d respectively w.r.t. the parameter α (first row) and β (second row).

depth–of–field appears infinite: The small hairs in the front as well as the compound eye are entirely sharp. The level of detail can also be seen regarding the depth map. Also here small structures such as hairs are clearly recognisable. For the depth map a grey value coding is used: The brighter the grey value, the larger the depth.
For the second real–world experiment, we apply our approach to a RGB focal stack capturing a coffee bean. The focal stack consists of 22 frames where 3 out of them can be seen in Figure 22(a). Also for this experiment, a coarse–to–fine strategy is used. This time we refrain from the variance method and use a constant depth value as initialisation on the coarsest grid. Furthermore, since the depth of the coffee bean only change smoothly, we apply Whittaker–Tikhonov regularisaion on d. The results are shown in Figure 22(b).

7 Conclusions

We have shown that modelling important physical properties such as the maximum–minimum principle helps to achieve a significantly better reconstruction quality in the depth–from–defocus problem. In contrast to many other papers in this field, we have employed robustification to cope with remaining imperfections in the imaging model, and we have made full usage of the colour information. Moreover, we have emphasised the benefits of a joint handling of denoising and depth–from–defocus over two separate models. By taking advantage of appropriate mathematics, we have mitigated the ill–posedness of the depth–from–defocus problem: On the one hand, the energy formulation allows for a stabilised model inversion via variational calculus. On the other hand, replacing the classical additive Euler–Lagrange formalism by its multiplicative variant preserves the positivity of the solution.

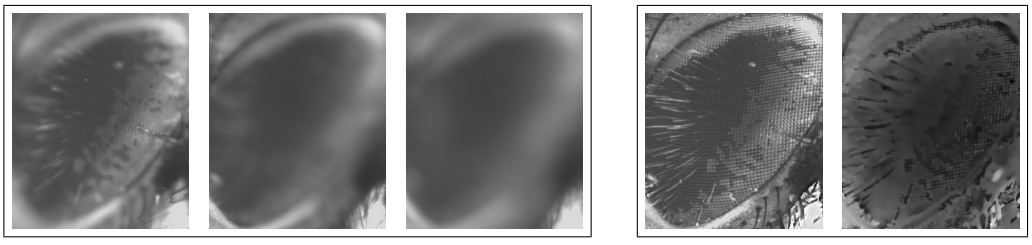

Figure 21: Focal stack of a house fly eye (grey scaled). This focal stack was provided by the Biomedical Imaging Group EPFL, Lausanne, Switzerland. **(a) Left box:** 3 out of 21 images of the focal stack. **(b) Right box:** Recovered pinhole image and depth profile ($\alpha = 25$).

Figure 22: Focal stack of a coffee bean. This stack was provided by the Computer Graphics Group, MPI for Informatics, Saarbrücken, Germany. **(a) Left box:** 3 out of 22 images of the focal stack. **(b) Right box:** Estimated pinhole image and reconstructed depth map ($\alpha = 20$).

Our work is an example how one can benefit from physically refined modelling in conjunction with multiplicative calculi. It is our hope that both concepts will receive more popularity in future computer vision models.

In our ongoing work we intend to incorporate further physical refinements into our model.

Acknowledgements. Our research has been partly funded by the *Deutsche Forschungsgemeinschaft (DFG)* through a *Gottfried Wilhelm Leibniz Prize* for Joachim Weickert and the Cluster of Excellence *Multimodal Computing and Interaction*.

References

[1] F. Aguet, D. Van De Ville, and M. Unser. Model–based 2.5–D deconvolution for extended depth of field in brightfield microscopy. *IEEE Transactions on Image Processing*, 17(7):1144–1153, 2008.

[2] Stephen W. Bailey, Jose I. Echevarria, Bobby Bodenheimer, and Diego Gutierrez. Fast depth from defocus from focal stacks. *The Visual Computer*, 31:1–12, 2014.

[3] B. A. Barsky and T. J. Kosloff. Algorithms for rendering depth of field effects in computer graphics. In *Proc. WSEAS International Conference on Computers*, pages 999–1010, Heraklion, Greece, July 2008. World Scientific and Engineering Academy and Society.

[4] M. Bertero, T. A. Poggio, and V. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, August 1988.

[5] S. Bhasin and S. Chaudhuri. Depth from defocus in presence of partial self occlusion. In *Proc. IEEE International Conference on Computer Vision*, volume 1, pages 488–493, Vancouver, Canada, July 2001.

[6] Max Born and Emil. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Pergamon Press, Oxford, 4th edition, 1970.

[7] R. Cant and C. Langensieoen. Creating depth of field effects without multiple samples. In *Proc. IEEE International Conference on Computer Modelling and Simulation*, pages 159–164, Cambridge, UK, March 2012.

[8] T. F. Chan and C. K. Wong. Total variation blind deconvolution. *IEEE Transactions on Image Processing*, 7:370–375, 1998.

[9] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge–preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2):298–311, 1997.

[10] S. Chaudhuri and A.N. Rajagopalan. *Depth From Defocus: A Real Aperture Imaging Approach*. Springer, Berlin, 1999.

[11] R. L. Cook, T. Porter, and L. Carpenter. Distributed ray tracing. In *Computer Graphics*, SIGGRAPH '84, pages 137–145, Minneapolis, USA, July 1984. ACM.

[12] P. Favaro, S. Osher, S. Soatto, and L. Vese. 3D shape from anisotropic diffusion. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pages 179–186, Madison, USA, June 2003.

[13] P. Favaro and S. Soatto. Shape and radiance estimation from the information divergence of blurred images. In D. Vernon, editor, *Computer Vision – ECCV 2000*, volume 1842 of *Lecture Notes in Computer Science*, pages 755–768. Springer, Berlin, 2000.

[14] P. Favaro, S. Soatto, M. Burger, and S.J. Osher. Shape from defocus via diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):518–531, 2008.

[15] I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Dover, New York, 2000.

[16] L. Hong, J. Yu, C. Hong, and W. Sui. Depth estimation from defocus images based on oriented heat–flows. In *Proc. IEEE International Conference on Machine Vision*, pages 212–215, Dubai, UAE, 2009.

[17] Hao Hu and Gerard de Haan. Adaptive image restoration based on local robust blur estimation. In Jacques Blanc-Talon, Wilfried Philips, Dan Popescu, and Paul Scheunders, editors, *Advanced Concepts for Intelligent Vision Systems*, volume 4678 of *Lecture Notes in Computer Science*, pages 461–472. Springer, 2007.

[18] P. J. Huber. *Robust Statistics*, volume 1. Wiley, Chichester, 2004.

[19] H. Jin and P. Favaro. A variational approach to shape from defocus. In A. Heyden, G. Sparr, M. Nielsen, and P. Johansen, editors, *Computer Vision – ECCV 2002*, volume 2351 of *Lecture Notes in Computer Science*, pages 18–30. Springer, Berlin, 2002.

[20] Nayar S. K. and Nakagawa Y. Shape from focus. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(8):824–831, Aug 1994.

[21] Cosmin Ludusan and Olivier Lavialle. Multifocus image fusion and denoising: A variational approach. *Pattern Recognition Letters*, 33(10):1388 – 1396, 2012.

[22] David Luenberger and Ye. *Linear and Nonlinear Programming*. Springer, New York, 3rd edition, 2008.

[23] A. Marquina and S. Osher. A new time dependent model based on level set motion for nonlinear deblurring and noise removal. In M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, editors, *Scale-Space Theories in Computer Vision*, volume 1682 of *Lecture Notes in Computer Science*, pages 429–434. Springer, Berlin, 1999.

[24] Vinay P. Namboodiri and Subhasis Chaudhuri. Use of linear diffusion in depth estimation based on defocus cue. In *Proc. Indian Conference on Computer Vision, Graphics and Image Processing*, pages 133–138, Kolkata, India, December 2004. Allied Publishers.

[25] Vinay P. Namboodiri and Subhasis Chaudhuri. On defocus, diffusion and depth estimation. *Pattern Recognition Letters*, 28(3):311–319, 2007.

[26] Vinay P. Namboodiri, Subhasis. Chaudhuri, and S. Hadap. Regularized depth from defocus. In *Proc. IEEE International Conference on Image Processing*, pages 1520–1523, San Diego, USA, October 2008.

[27] S. Osher and L. Rudin. Total variation based image restoration with free local constraints. In *Proc. IEEE International Conference on Image Processing*, volume 3, pages 31–35, Austin, Texas, 1994.

[28] A. P. Pentland. A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(4):523–531, 1987.

[29] N. Persch, C. Schroers, S. Setzer, and J. Weickert. Introducing more physics into variational depth-from-defocus. In X. Jiang, J. Hornegger, and R. Koch, editors, *Pattern Recognition*, volume 8753 of *Lecture Notes in Computer Science*, pages 15–27. Springer, Berlin, 2014.

[30] Said Pertuz, Domenec Puig, and Miguel Angel García. Analysis of focus measure operators for shape-from-focus. *Pattern Recognition*, 46:1415–1432, 2013.

[31] Matt Pharr and Greg Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann, San Francisco, 2004.

[32] A.N. Rajagopalan and S. Chaudhuri. A variational approach to recovering depth from defocused images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(10):1158–1164, Oct 1997.

[33] Przemyslaw Rokita. Fast generation of depth of field effects in computer graphics. *Computers & Graphics*, 17(5):593 – 595, 1993.

[34] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.

[35] Murali Subbarao. Parallel depth recovery by changing camera parameters. In *Proc. IEEE International Conference on Computer Vision*, pages 149–155, Washington, USA, December 1988.

[36] Murali Subbarao and Gopal Surya. Depth from defocus: A spatial domain approach. *International Journal of Computer Vision*, 13:271–294, 1994.

[37] Murali Subbarao and Jenn-Kwei Tyan. Noise sensitivity analysis of depth-from-defocus by a spatial-domain approach. In *Proceedings of SPIE 3174*, pages 174–187, San Diego, USA, July 1997.

[38] Satoshi A. Sugimoto and Yoshiki Ichioka. Digital composition of images with increased depth of focus considering depth information. *Applied Optics*, 24(14):2076–2080, 1985.

[39] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady*, 4:1035–1038, 1963.

[40] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[41] Y. Wei, Zaili Dong, and Chengdong Wu. Global depth from defocus with fixed camera parameters. In *Proc. IEEE International Conference on Mechatronics and Automation*, pages 1887–1892, Changchun, China, August 2009.

[42] M. Welk and J. Nagy. Variational deconvolution of multi–channel images with inequality constraints. In J. Martí, J. Benedí, A. Mendonça, and J. Serrat, editors, *Pattern Recognition and Image Analysis*, volume 4477 of *Lecture Notes in Computer Science*, pages 386–393. Springer, Berlin, 2007.

[43] M. Welk, D. Theis, and J. Weickert. Variational deblurring of images with uncertain and spatially variant blurs. In W. Kropatsch, R. Sablatnig, and A. Hanbury, editors, *Pattern Recognition*, volume 3663 of *Lecture Notes in Computer Science*, pages 485–492. Springer, Berlin, 2005.

[44] E. T. Whittaker. A new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:65–75, 1923.

[45] Y.-L. You and M. Kaveh. Anisotropic blind image restoration. In *Proc. IEEE International Conference on Image Processing*, volume 2, pages 461–464, Lausanne, Switzerland, Sep 1996.

[46] Y.-L. You and M. Kaveh. A regularization approach to joint blur identification and image restoration. *IEEE Transactions on Image Processing*, 5(3):416–428, Mar 1996.