Computers & Graphics 74 (2018) 109–118

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Automated outdoor depth-map generation and alignment☆

Martin Čadík[a,*], Daniel Sýkora[b], Sungkil Lee[c,*]

[a] CPhoto FIT, Faculty of Information Technology, Brno University of Technology, Czech Republic
[b] Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic
[c] College of Software, Sungkyunkwan University, Republic of Korea

ARTICLE INFO

Article history:
Received 18 January 2018
Revised 23 March 2018
Accepted 1 May 2018
Available online 16 May 2018

Keywords:
Image enhancement
Synthetic depth
3D terrain
Free-form warping
Image registration
Synthetic camera

ABSTRACT

Image enhancement tasks can highly benefit from depth information, but the direct estimation of outdoor depth maps is difficult due to vast object distances. This paper presents a fully automatic framework for model-based generation of outdoor depth maps and its applications to image enhancements. We leverage 3D terrain models and camera pose estimation techniques to render approximate depth maps without resorting to manual alignment. Potential local misalignments, resulting from insufficient model details and rough registrations, are eliminated with our novel free-form warping. We first align synthetic depth edges with photo edges using the as-rigid-as-possible image registration and further refine the shape of the edges using the tight trimap-based alpha matting. The resulting synthetic depth maps are accurate, calibrated in the absolute distance. We demonstrate their benefit in image enhancement techniques including reblurring, depth-of-field simulation, haze removal, and guided texture synthesis.

1. Introduction

A limited configuration in taking photographs does not always lead to the highest quality, and often motivates enhancement of photographs. Computational photography has addressed such limitations, which introduces additional flexibility on focus, exposure, and depth [1–3]. Among them, depth information, on which we focus here, can greatly facilitate diverse image manipulations, such as refocusing, dehazing, texture synthesis, and image editing [4–7].

Outdoor photographs (e.g., natural landscapes) represent by far the biggest group in many media services [8], but their direct depth estimation poses a challenge. They are usually monocular, which has a low chance to work with typical structure-from-motion. Within-image features, such as airlights or textures [9,10], help, but are not always available. Range sensors [11] are applicable only to limited distance ranges (only up to tens of meters).

One better alternative can be an indirect estimation from a 3D terrain model, which renders the reference depth map as previously suggested by Kopf et al. [12]. The terrain model is already available for the whole planet and recent photographs are usually geo-tagged (e.g., the global positioning system; GPS), which can serve as a strong external cue. Further, this approach can be distinguished for its higher resolution and accuracy; the direct estimation may yield a coarser resolution or wrong outcomes.

However, the 3D terrain model may be insufficient in its resolution and details (e.g., textures and objects). Also, a precise registration between real and virtual views is challenging, where the alignment errors result in visible artifacts in edited images. Manual registration starting at a rough initial guess can help [13], but is laborious and inappropriate for massive batch processing.

In this work, we present a *fully automatic* framework for depth-map generation and alignment for an outdoor photograph. A virtual camera is first localized with the geo-tagging information of a photo and recent camera pose estimation techniques. Then, the 3D terrain model is rendered at the virtual camera to produce an initial approximate depth map. Local inaccuracies, resulting from the rough registration and insufficient details of the model, are subsequently reduced using our novel *automatic* free-form warping. We first align discontinuities in the synthetic depths with photo edges using the as-rigid-as-possible image registration. The shape of the edges is further refined using the tight trimap-based alpha matting. The resulting depth map, synthesized from the geo-referenced terrain model, is *absolute* (calibrated in meters). We show its benefit in image enhancement tasks, including refocusing, defocus manipulation (see Fig. 1), dehazing, and texture synthesis.

☆ This article was recommended for publication by Dr M Kim.

* Corresponding authors.
E-mail addresses: cadik@fit.vutbr.cz (M. Čadík), sungkil@skku.edu (S. Lee).

https://doi.org/10.1016/j.cag.2018.05.001
0097-8493/

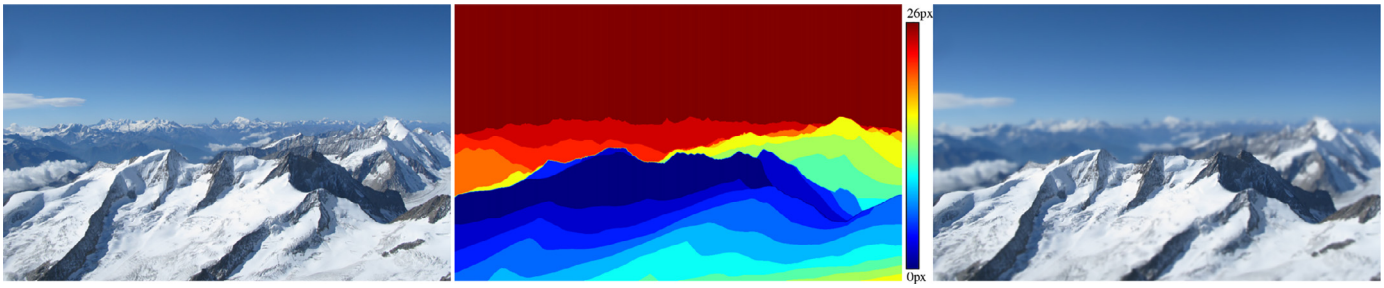

Fig. 1. Transforming an outdoor photograph into a model-like look. An automatically generated synthetic depth map is used to calculate plausible blur kernel size map (middle) to simulate shallow depth-of-field (right) in landscape images (left), where such an effect cannot be achieved using standard optics for physical reasons. Virtual camera: full-frame, f-number=1.0, focal length=1200 mm, focus distance=5 km.

2. Related work

We briefly review previous work on depth map reconstruction and its major applications including defocus manipulation and dehazing.

2.1. Depth map reconstruction

Robust depth map reconstruction is an ongoing subject of interest. A typical approach is to rely on stereo image pairs [18,19] or multiple/multiview images [20–22]. More recently, short-distance range-sensing devices, such as Kinect [11], improved the availability of depth maps in indoor environments [23].

Multiple images are not easily available in practice, and single-image processing has also been intensively studied. Semi-automatic user interaction often helps [24–26], and further modification to hardware or light patterns proved its utility, such as light fields [27–31], coded aperture [32] or structured light method [33].

Depth maps can further be generated semi-automatically using sparse samples (seeds) provided by the user [34–37]. In these approaches, anisotropic diffusion is used to propagate depth information from seeds to the rest of the image. The key assumption here is that the gradient of the resulting depth map should roughly correspond to the color gradient.

The previous methods are not applicable to ours which is targeting on *single outdoor* photographs. The stereo vision techniques require multiple images, while the range sensors work only for a small range of distances. Computational photography requires special hardware or modification to the aperture. For the diffusion-based techniques, real photos often violate their main assumption about compatible depth and color gradients, and also, the positions of depth seeds require to be accurate. Otherwise, the diffusion will propagate small misalignments to a notably larger area and lead to notable artifacts.

Kopf et al. showed the combination of geo-tagging and 3D models can be used for fairly accurate geo-registration and many applications including dehazing and relighting [12]. The geo-tagging already allows us to select an effective subset for structure from motion [38], but when combining with the available 3D models, we can directly render a depth map. The combination even enables to assign geo-locations and labels onto pixels [39], point clouds [38,40], or annotate photos [41]. Nevertheless, the depth map is not pixel-perfect and requires fine alignment; we address this issue in the present work.

While the majority of previous approaches estimate *relative* depth maps, our solution can generate *absolute* depth map from the geo-referenced digital terrain model. This is highly beneficial in many image enhancement applications; for instance, there are more chances in estimating kernels for defocus blur or other effects.

Similarly to our goals, the method proposed by Kopf et al. [12] allows to synthesize absolute depth maps. However, it requires a *user-assisted* interaction for registration, which motivates for our novel depth free-form warping step (Section 3).

2.2. Defocus manipulation

The defocus blur, caused by shallow depth of field (DOF), is pronounced in indoor photos or films, but hardly exhibited in outdoor photography (even with large lenses). Its capture is inherently restricted to a particular configuration (e.g., focus and *f*-number). Thus, its post-reproduction for novel configurations (with computational photography) drew attention for refocusing [32,33] or defocus magnification [42].

Another mainstream is the postprocessing of a usual single-view image, which comprises deblur and refocusing. Typically, the deblur involves blind deconvolution using known priors to estimate kernels [24,43]. However, a precise solution requires to consider geometric visibilities similarly to the distributed ray tracing [44]. Our solution and its absolute depth information can facilitate the estimation of per-pixel local blur kernels to some extents, enabling non-blind deconvolution. Refocusing can also benefit from ours, which can use a precise rendering technique [45].

2.3. Dehazing

Outdoor photographs are often hazed by atmospheric scattering that can be characterized by medium *transmission maps*, depending heavily on depths. Most previous work has focused on depth estimation and radiance recovery, including Markov random fields (MRF) [46], independent component analysis [47], dark channel prior [48], factorial MRF [49–52], and machine-learning approaches using random forests [53,54], and convolutional neural networks [55]. In our case, geo-referenced absolute depth maps are used for dehazing, which can be more precise than the previous ones. We obtain such maps automatically without the previous manual image-to-model registration [12].

3. Automatic depth-map generation and alignment

In this section, we describe our fully-automatic approach to depth-map generation from a single color image as well as a technique used for the final depth map refinement (see Figs. 2 and 3 for summary).

Retrieval of 3D terrain model. A Google-Earth-like digital terrain models are currently available for the whole planet. They are acquired from satellites and/or planes and published in form of geo-referred digital elevation maps (DEMs) even for less accessible regions. Such models are sufficient for our purposes (i.e., *outdoor photographs*), however in general case, a 3D model might not be

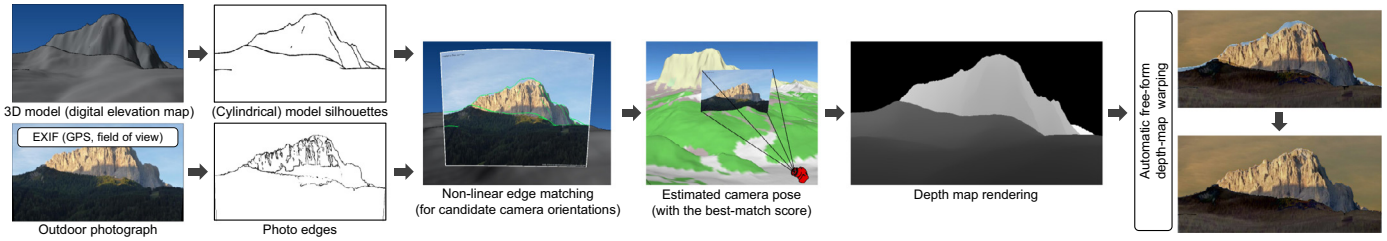

Fig. 2. Overview of our fully automatic depth-map generation framework from a single landscape photograph. Based on the EXIF information of the photograph, the camera pose to render the 3D terrain model is automatically aligned with the image. Then, the initial coarse depth map is rendered from the model using the estimated camera pose; some inaccuracies may show up due to insufficient precision of the model or due to camera alignment errors. The final depth map is refined using the free-form warping to match the local features of the input photograph.

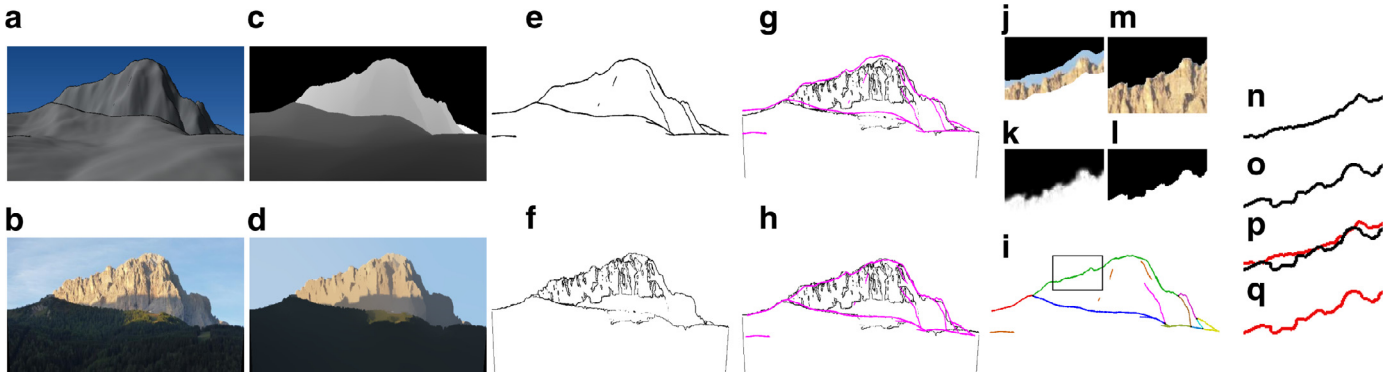

Fig. 3. Aligning model depth map with the input photograph—the 3D model (a) is roughly aligned with an input photo (b), the depth map (c) is extracted using the estimated camera location and an intrinsic image [14] of the input photo (d) is computed. Depth discontinuities are extracted in the depth map (e) and edges are detected in the intrinsic image (f). Initially, model edges are misaligned with respect to photo edges (g), to reduce this misalignment as-rigid-as-possible image registration [15] is used (h), then edges are subdivided into individual segments (i) and tight trimap is constructed for each segment (j), alpha matte [16] is computed (k) and thresholded (l) to get the final refined shape of the photo edge (m). Given the initial model edge (n) and the refined photo edge (o), deformable image registration [17] is used (p) to obtain the final sub-pixel accurate alignment (q).

available (e.g., for indoor environments). We experimented with the following publicly available DEM terrain datasets: the Alps (24 m spaced samples (px)[1]), National Elevation Dataset (8 m/px, USGS[2]), and Eastern Europe models (10 m/px[3] and 5 m/px[4]). In case of an areal overlap of the terrain models, we used the one with the highest available resolution.

In case of an outdoor scene, the DEM is generally sufficient for camera *localization* and *pose estimation*, and a synthetic depth map may then be *generated* easily. However, existing elevation models still do not capture trees, small buildings, cars, and other man-made features in a sufficient detail. For this reason, we propose to finally *refine* the generated *depth map* using the free-form registration with the input photograph.

Localization of Virtual Camera. To render the initial depth map, the location of a real camera (used to capture the outdoor photograph) requires to be known in advance. By default, we can locate the position of unknown camera using the structural features of an input photo. Specifically, we use skyline/contours and further geometric constraints as proposed by Baatz et al. [56], when the image lacks information about camera location. Other possibilities include cross-view image geolocalization [57], direct data-driven regression [58] or classification [59] of the camera location, image registration into the 3D structure acquired by structure from motion [40], image retrieval from a geo-referenced image database [60], and others [61]. However, they are exhaustive and inaccurate in some cases. Fortunately, approximate locations can be easily found from the picture itself in many cases; GPS is integrated with many recent cameras and smartphones and its information is recorded in EXIF tags. Hence, we assume the approximate location of the camera is already known and in the following steps we sample only near proximity of the known location.[5]

Camera pose estimation. Given the camera location, we automatically estimate its *pose*, i.e. all the unknown camera orientation angles (yaw, pitch, and roll). We implemented a visual camera orientation estimation in a way similar to that of Baboud et al. [41]. The method is based on matching the edges detected in the photograph with the synthetic silhouettes rendered from the terrain model.

More specifically, we first exploit the image EXIF data (assisted by a camera database) to perform rectilinear projection with the known field-of-view. We then render model silhouettes as depth discontinuities into a 2D cylindrical image, which is vectorized into a silhouette edge map. This map is then aligned with the image edges by means of vector cross-correlation followed by a non-linear matching metric [41] (see Fig. 2).

For the image edge detection, we developed a novel weighted edge estimator using the learning-based framework [62]. During matching process described above, the edges are thus assigned relative importance according their visual appearance. For finding salient silhouette edges, we predict a 16×16 px edge map from a larger 32×32 px image patch. Individual predictions are averaged to produce a soft edge map for the whole input image. The

[1] http://www.viewfinderpanoramas.org.
[2] http://ned.usgs.gov.
[3] http://www.geoportal.sk.
[4] http://geoportal.cuzk.cz.

[5] The area is sampled uniformly as a grid of 9×9 samples with 0.001° resolution in both N-S and W-E directions.

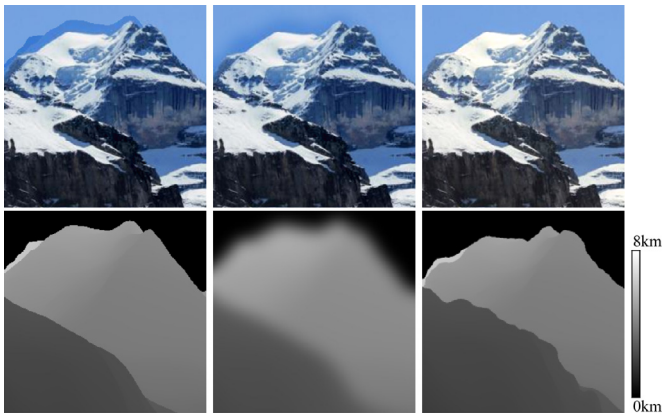

Fig. 4. Dehazing artifacts due to coarse synthetic depth map (left) can be partially mitigated by blurring the depth map (middle), and fully removed using the proposed free-form depth warping (right).

learning problem is solved using structured random forests. In order to use standard node splitting criteria, the structured space of labels $\mathcal{Y}$ is mapped to a discrete set of labels $\mathcal{C}$ by a two-stage mapping via an intermediate space $\mathcal{Z}$ at each node. The learning-based framework [62] assumes segmentation maps being available for training. Instead, we use pre-rendered depth maps. To be able to use depth maps as labels, we redefine the intermediate mapping $\Pi : \mathcal{Y} \rightarrow \mathcal{Z}$ to produce a vector that encodes depth difference $y(j_1) - y(j_2)$ for every unique pair of indices $j_1 \neq j_2$ within a label patch $y \in \mathcal{Y}$. In practice, we sample $m = 256$ dimensions of $\mathcal{Z}$, resulting in a node-specific reduced mapping Π_Φ, which is then further discretized as in the original paper.

The camera pose estimation procedure is repeated for each candidate in the proximity of the approximate camera location, and the result with the highest matching score is selected for further processing.

Depth map rendering. Given the camera parameters estimated in the steps described above, we can easily render the depth map from the terrain model (in our case, using ray casting). The obtained depths calibrated in meters are finally stored in an image file of high-dynamic-range format (to facilitate the absolute distance estimation).

However, the synthesized depths may exhibit local misalignment between the depth map and the photo, resulting from the coarse resolution of the 3D model, unknown non-rigid projection parameters of the photos, or other inaccuracies in the alignment process (e.g., occlusion of silhouette edges). This problem may be alleviated by blurring the depth map (Fig. 4), but for better registration, we propose *free-form depth map warping* in what follows.

Automatic free-form warping of depth map. To register the model's depth map with the input photo (Fig. 2), we propose a novel automatic free-form warping solution that resolves two key challenges: (1) cross-domain registration between the color image and the depth map and (2) potentially large misalignments.

For the first challenge, we unify the registration domains by extracting dominant discontinuities in the depth image (Fig. 3c) as well as the photo (Fig. 3b). The model's depth map expresses the discontinuities with *model edges* (Fig. 3e), which is extracted by hard-thresholding the magnitude of a gradient field of the depth map. Corresponding *photo edges* are computed from the intrinsic image [14] (Fig. 3d), which removes spurious edges caused by shading. Then, we again extract the gradient magnitude in the color domain with thresholding (Fig. 3f).

To cope with the second challenge (the large mismatches between model and photo edges), we use a previous iterative as-rigid-as-possible deformation method [15] (Fig. 3h). Its key advantage here is its block matching that allows us to find a globally optimal registration on a small neighborhood, while as-rigid-as-possible regularization suppresses the excessive deformation of edges.

Once the model edges (Fig. 3j) are roughly aligned around the photo edges, we further refine the shape of the photo edges using a trimap-based matting. To do so, we detect junctions and end points on the model's edge map and subdivide edges into individual segments. Edges of each segment are eroded to build the trimap by setting three distinct regions: apparent background (black), apparent foreground (white), and unknowns. Then, we compute an alpha matte (Fig. 3k) using a closed-form approach [16]. The matte is thresholded around 0.5 (Fig. 3l) to refine the shape of the photo edges (Fig. 3m).

Finally, we warp the depth map by aligning the roughly aligned model edges to the refined photo edges. We perform the sub-pixel accurate deformable registration [17] with the input photo (Fig. 3q). Then, the resulting deformation field is used to warp the initial depth map to the final depth map (Fig. 2).

4. Experimental results

Qualitative evaluation. Our system produces *absolute* depth maps in high quality; see Fig. 5 (right). Unlike ours, the previous methods operating directly on image pixels [48,63] may result in *relative* and *noisy* distances; Fig. 5 (middle). The depth map quality comparable to our rigid alignment can be achieved with the manual model-to-photo alignment [12], but our results can be produced automatically and in a shorter time, as discussed below. Moreover, the manual alignment may also benefit from our free-form warping, because the input terrain model is hardly perfect in terms of object details.

The spatial accuracy of the resulting depth maps depends on the resolution of the terrain model and on the distances captured in the photo (i.e., the visual angle subtended by a pixel). Hence, the farther the area to the camera, the higher quality achieved. Fig. 6 shows depth map edges overlaid on the photo with different resolutions of the DEMs. Our experiences show that currently available DEMs ($<$ 8 m spaced samples) constitute sufficient resolutions for objects farther than 500 m from the camera. This issue can be partially mitigated by blurring the depth map, however this, in general, motivates for our free-form warping; see Fig. 4.

Finally, we compared our free-form depth warping with DeepMatching [64], a recent correspondence matching method (to facilitate further warping); see Fig. 7. DeepMatching uses multi-level correlation architecture to handle non-rigid deformations and to determine dense correspondences. Our warping well aligns synthetic depth edges with the details captured in the photo resulting in correct depth map (Fig. 7, right column). In contrast, DeepMatching (Fig. 7, middle column) fails to establish relevant correspondences between the depth map and photo due to their different modalities.

Quantitative evaluation. Quantitative evaluation of our free-form depth map warping is difficult, because there exist no ground-truth datasets (i.e., accurate depth maps for real natural landscape photos). Instead, we used *manually-aligned* synthetic depth maps as *reference* depth maps for comparison. The depth values are assigned by our technique for depth map generation, but the depth maps are aligned manually.[6]

The evaluation used three photos and their reference depth maps. For each photo, we then generated 1000 depth maps (inputs to the automatic warping) randomly distorted in yaw, pith,

[6] The new dataset is available for download at: http://cphoto.fit.vutbr.cz/depth/.

Fig. 5. Given a single color input image, our system creates more plausible absolute depth maps (right) unlike previous depth map estimations (e.g., dark channel prior [48], or deep learning based method [63]; middle).

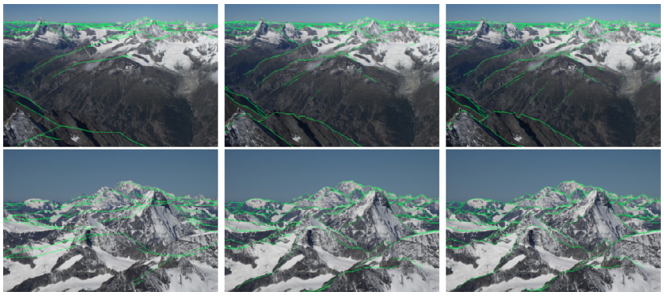

Fig. 6. The effect of DEM resolution on the spatial quality of the depth map. DEM resolutions of 480 m, 48 m, and 24 m (left to right).

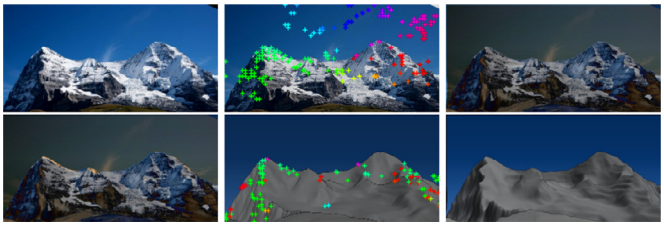

Fig. 7. Comparison of DeepMatching algorithm [64] (middle column) with our free-form warping (right column). Left column: input photograph and original misaligned depth map. While ours well aligns edges which results in a correctly warped depth map, DeepMatching fails to establish dense correspondences making warping impossible (colored crosses denote correspondences).

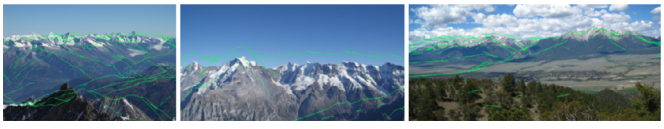

Fig. 8. Experimental evaluation set-up. In reading order: test images 1–3 with example distortions in camera yaw ($\alpha = +0.05$rad), pith ($\beta = +0.05$rad), and roll ($\gamma = +0.05$rad).

and roll; see Fig. 8. Each rotation angle was altered by an independent random additive coefficient drawn from normal distribution (standard deviation σ=0.01rad). The input depth maps with the original photo were fed into our free-form warping algorithm to produce output depth maps.

We compared the accuracy of outputs of our depth map warping against the recent single-image depth-map synthesis techniques, monodepth [63] and dark-channel-prior methods [48];

Table 1
Quantitative evaluation of depth map accuracy (E_{absr}).

Method	Image1	Image2	Image3	Average
Monodepth [63]	2.647	0.476	2.247	1.790
Dark-channel-prior [48]	9.462	0.601	6.932	5.665
Our method	**0.335**	**0.026**	**0.134**	**0.165**

there exists no competitive *automatic depth-map warping* method. As the other two methods give only *relative depth values*, we normalized all the predicted depths (including ours and ground-truths) linearly to the interval of [0,1] for the following evaluation. For each method, difference with the reference depth maps is quantified using commonly-used *abs relative difference* [66]: $E_{\text{absr}} = \frac{1}{|T|}\sum_{i\in T}|d_i - d_i^*|/d_i^*$, where T is the resolution of input image, d and d^* are the estimated depth and reference depth, respectively. Table 1 shows errors (lower is better). As shown in the table, the errors of our depth map warping are order of magnitude better than the other two methods.

Moreover, we also quantify the robustness of the proposed free-form warping to particular errors in camera pose estimation. Fig. 9 shows alignment errors with regard to camera misalignment specified by angular distortions. The sensitivity of our method to errors in yaw, pith, and roll is similar; i.e., all three camera rotations affect the error to the same degree. To test this statistically, we use 3-way analysis of variance (ANOVA) [67], which includes factors of yaw, pitch, and roll. The null hypothesis is "there is no significant difference in warping error between camera orientation distortions in yaw, pitch, and roll." The ANOVA results showed that the null hypothesis is not rejected ($F_{yaw}(1, 2999) = 0.88$, $p_{yaw} = 0.81$; $F_{pitch}(1, 2999) = 1.22$, $p_{pitch} = 0.18$; $F_{roll}(1, 2999) = 0.75$, $p_{roll} = 0.86$).

Performance. Our unoptimized C++ implementation (on Intel i7 2.4 GHz, 8 GB memory) requires on average 1 and 4 min. for image-to-model alignment and free-form warping, respectively; depth map rendering is marginal. The free-form warping spends most of time on the L1 intrinsic decomposition. This can be accelerated using a better edge detector (e.g., data-driven method [62]), which is a good avenue for future work.

For comparison, we implemented the manual alignment in a way similar to Kopf et al. [12]. Our experience is in accord with previous measurements [13]; the interactive session requires on average 3 min. for a skilled user to align the photo with the model.

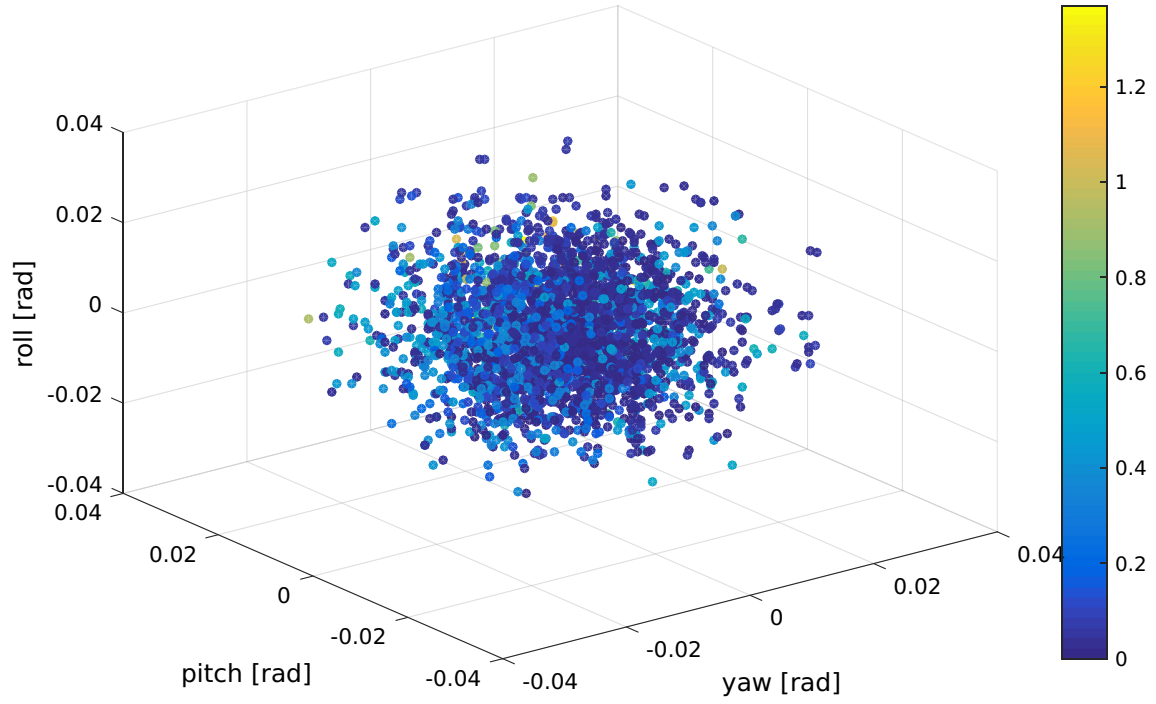

Fig. 9. Depth map alignment error (E_{absr}) for our free-form warping given distortions in camera yaw, pith, and roll.

Note that a final depth refinement (i.e., our free-form warping) is still required to reduce misalignments from insufficient details of the terrain model against the photo.

5. Applications of absolute depth maps

This section demonstrates several applications that profit from our automatic depth map generation. In particular, we present a novel image refocusing/defocusing algorithm that benefits from the known *absolute* distances stored in the depth map. We also show further benefits of our approach in dehazing, and guided texture synthesis.

5.1. Image refocusing and defocusing

The image refocusing/defocusing algorithm we propose works in three steps. First, the unknown focal plane is estimated from detected focused pixels whose absolute distances are given in our depth map. Then, space-variant blur kernels are calculated to refocus the input photograph using non-blind deconvolution. Finally, the depth-of-field effects are simulated via post-processing or image-based ray-tracing. Please note that both the focal plane estimation and space-variant deconvolution steps are feasible thanks to the absolute depth estimated in Section. 3.

Estimation of focal plane distance. The distance to the focal plane from the camera, usually denoted as *point of focus* (PoF) in photography, is crucial for subsequent refocusing steps. Since it is difficult to precisely measure the focal plane distance without knowing the accurate capture conditions, we instead estimate the focal plane distance using a *focus measure*.

The idea of our focus measure is to select a *median* depth of focused pixels in the image, which is a reasonable approximation to the true focal distance. We detect (sharply) focused pixels using Laplacian of Gaussian (LoG), which performs well in shape-from-focus evaluations [68]. We apply thresholding to the LoG response of the image, and keep only the sharpest pixels (i.e., top 2%). We then query the synthetic depth map for the *absolute distances* of the detected focused pixels, and use their median depth as the focal plane distance.

Refocusing by space-variant non-blind deconvolution. While outdoor photographs are less affected by the defocusing, some of them might still exhibit slight defocus blur. In such a case, a refocusing/deblur is required to recover the sharpness of the photographs, resulting in all-focused images suitable for further processing.

In general, the deconvolution kernel (point-spread function, PSF) is unknown (so, we call *blind* deconvolution), and multiple kernels or images can produce the same output. Thus, the blind deconvolution for refocusing, in particular with a single image, is an ill-posed and challenging problem. Previous single image-based methods simultaneously assessed the spatially-varying blur kernel and the sharp image [69–72].

Depth information available from our absolute depth maps is greatly helpful in enhancing the accuracy of deconvolution or facilitating *non-blind* deconvolution. Assuming the image blur comes solely from the defocus (i.e., no motion blur), we can *calculate* the extents of the spatially-varying PSFs, as it varies depending on the depth and the focal distance (estimated in the previous step).

Unlike general (motion) blur, the defocus blur does not have much variation in the shape of PSF. The PSF generally resembles the shape of the aperture, and the parameters are often available from EXIF information (f-number or aperture stop, focal length, and sensor size). Assuming diffraction-free and aberration-free imaging, the PSF can be a simple circular or polygonal shape with the constant/Gaussian intensity profile. This information allows us to approximate the spatially-variant PSFs, as illustrated in Figs. 1 and 10. For text brevity, the formulae to calculate the spatially-variant PFSs are given in Appendix A.

Having the approximate PSFs for the image, we proceed with the non-blind deconvolution. Importantly, the deblurring algorithm must accommodate PSFs with discontinuities, because the depth may vary significantly in outdoor images. To this end, we implemented a space-variant deblurring method based on constrained

Fig. 10. Simulation of shallow depth-of-field effects. The input image (first) has been artificially defocused using synthetic depth map, focusing on 6 km, 30 km, and 70 km (in the reading order). The virtual full-frame DSLR camera used f-number=1.0, and focal length=1200 mm.

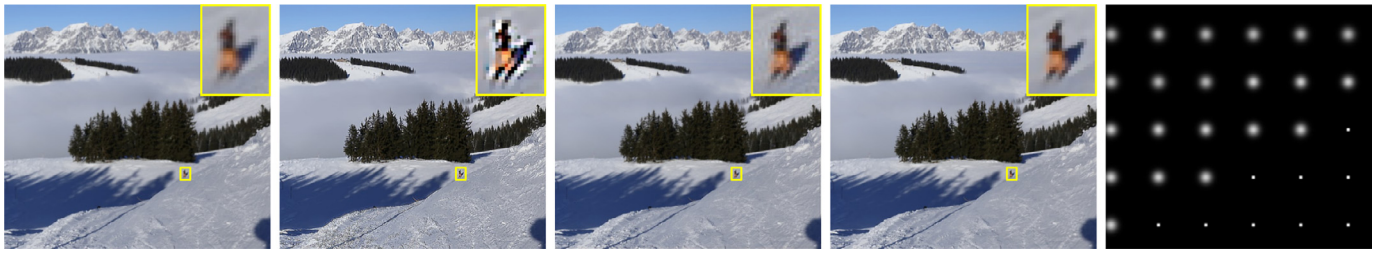

Fig. 11. Image refocusing results (best viewed in electronic version). In the reading order: original image, deblurring using regularized intensity [65], maximum likelihood blind deconvolution using our largest estimated blur kernel as a prior, and the proposed space-variant deblurring result. Right: an illustration of the space-variant kernels. The deconvolution techniques, that assume only single kernel, oversharpen the foreground. Our method refocuses the image adaptively and does not suffer from this problem.

Fig. 12. Comparison of ray-traced defocus blur (b–d) and Gaussian blur (f–h) generated from the color (a) and depth inputs (e). The black-edged boxes show blur kernel sizes (in pixels). The ray tracing better handles discrete depth boundaries than Gaussian blur does (see the red insets).

least square method with total variation regularization [73]. In contrast to previous approaches which resorted to estimations of PSFs and/or assumed only a uniform PSF, we *directly calculate* the space-variant kernels using absolute distances encoded in the depth map. This improves image refocusing results, as shown in Fig. 11. The full refocusing process requires on average 1.5 min. for 1M pixel image (on Intel i7, 2.4 GHz, 8 GB memory).

Defocus manipulation and depth-of-field simulation. Finally, we can re-blur the refocused sharp image as the user wishes. Besides faithful reproduction of the depth-of-field (DoF) to real cameras, shallower DoFs can be expressed to enhance saliency of objects, which would normally be impossible to capture in real ones (Figs. 10 and 12).

An accurate depth map, which we acquire in Section. 3, is utilized again for an adaptive kernel. Similarly to the refocusing, the shape of the PSF is given by the depth map, point of focus, and parameters of, this time, the *simulated camera*; see Appendix A for details.

We first experimented with a separable filtering with Gaussian-kernel, but such a simple convolution-based blur often fails around discontinuous depth boundaries (Fig. 12). The depth transition can be better handled by incorporating precise geometric visibility, which can be derived from the depth information. For this purpose, we also implemented the state-of-the-art GPU-based ray tracer [45], which can blend multiview images precisely and deliver better quality (Fig. 12). We implemented both methods using GPU (on Intel i7 2.4 GHz with NVIDIA GTX 980 Ti and OpenGL), and they performed in real time (e.g., 1–2 ms for 1024×680 resolution).

5.2. Single image haze removal

Natural landscape images are often degraded by haze due to atmospheric absorption and scattering. *Image dehazing* for removing such phenomena is generally a challenging problem, because the amount of haze depends on the distance from the camera. The main effort of single image dehazing algorithms [48,51] was directed to depth estimation. In our case, however, the *absolute* depth is given in an accurate depth map and we can directly proceed to the recovery of the scene radiance. Having solved the main dehazing issue implicitly via the estimated depth map, we proceed

Fig. 13. Effect of the automatically estimated dehazing parameter β. In the reading order: $\beta = 0, 1.5\times10^{-6}$ and 3.5×10^{-6}.

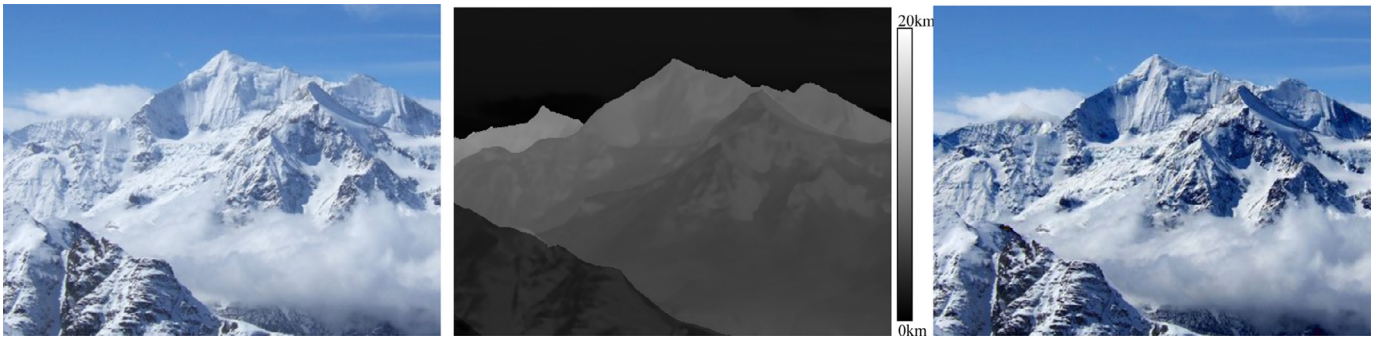

Fig. 14. Example of the single-image haze removal (right) for the input photo (left) using synthetic depth map (middle). Notice the spiky peak on the horizon, which has been completely obscured by clouds in the input photo.

similarly to He et al. [12,48]. We recover the scene radiance on per-pixel basis as follows: $J_x = A + (I_x - A)/(\max(t_x, t_0))$, where J_x is the recovered radiance in the pixel x, A is the global atmospheric light, I_x is the observed pixel irradiance, t_x is the medium transmission given by the depth map, and $t_0 = 0.1$. More specifically, $t_x = e^{-\beta d}$, where d is the depth of the point x, and β is the scattering coefficient. We set β so that $t_x = 0.75$ for the furthest point in the photo, i.e. $\beta = 0.3/d$, which leads to good results (Fig. 13). The coefficient A is initialized automatically with the average color of the 0.1% brightest pixels from the dark channel [48] in the sky area. The dehazing results are obtained at interactive speeds and they are shown in Fig. 14.

5.3. Guided texture synthesis

Guided texture synthesis (texture-by-numbers) is a variant of popular image analogies framework [74]. It allows us to transfer a texture from a given exemplar to a target image using guiding feature maps. In our scenario, *range data* (depth map) and *synthetic shading* can be used to guide semantically meaningful transfer of texture from the existing photograph to a virtual scene (Fig. 15). Note how the corresponding values in the range map and shading image help the algorithm to synthesize proper texture at particular locations, e.g., snowcapped peaks or shadows in lowlands. To implement this, we used *StyLit* method [75] (current state-of-the-art in guided texture synthesis). However, we replaced LPE-based guiding channels used in StyLit with our depth map and shading and run the synthesis algorithm, which resulted in faithful synthetic images as shown in Fig. 15. The advantage of StyLit as compared to the original greedy approach of Hertzmann et al. [74] is that it performs texture optimization which jointly satisfies texture coherence as well as matching of guiding channels. In addition to that it also adaptively encourages uniform usage of source texture patches which significantly improves the overall quality of the synthesis.

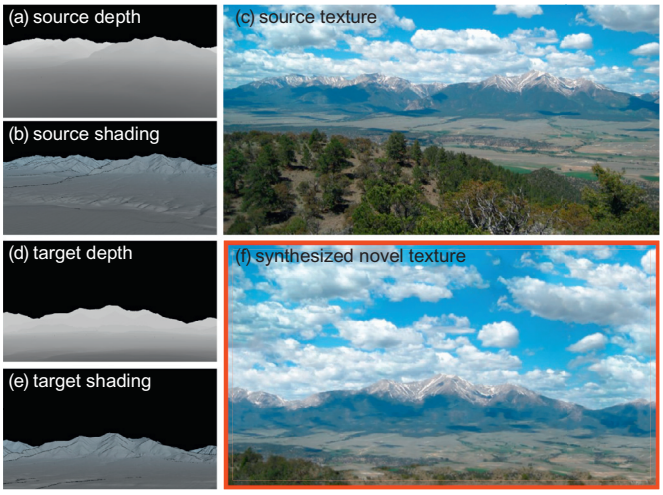

Fig. 15. Example of guided texture synthesis. From the source texture, its depth map and shading, we automatically synthesize a novel texture (the red box) for the target depth and shading. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusions and discussions

In this paper, we proposed an automatic approach to acquire depth maps of natural landscape images. The absolute depth is rendered from a digital elevation model, that is automatically aligned with the input photograph. To match the tiny details of the photograph, which are not necessarily captured in the model, we proposed a free-form warping step. In this way, we obtained accurate depth maps calibrated in absolute distances. We showed this was beneficial in image editing and enhancement, in particular for refocusing and defocus manipulation. We further showed the benefit of our synthetic depths in dehazing, and guided texture synthesis.

Fig. 16. Foreground objects, which are not depicted in the synthetic depth map, may adversely affect the results of our algorithms (left: input image, right: biased dehazing result).

Limitations and future work. The main limitation of our approach resides in large foreground objects, which are often captured in the photograph but not in the model. The free-form warping step cannot cope with this case and artifacts may show up in subsequent image processing (Fig. 16). This could be potentially alleviated by foreground object classifiers, which may direct further research.

The depth map rendering step assumes a correctly estimated camera pose, i.e. camera location and orientation. This estimation is rather stable and efficient, when the approximate position of the camera is known (e.g. given the GPS reading stored as an EXIF tag). However, when camera positions are completely unknown, the pose estimation is much less reliable due to a large-scale exhaustive search. In that case, our pipeline resorts to purely visual camera geo-localization [56]. This is an extremely difficult task especially in outdoor environments, and as such, it is a topic of intensive ongoing computer vision research.

Our as-rigid-as-possible deformation method relies on the existence of sufficiently strong gradients to align mismatches between the model and photo edges in the input photograph. This assumption may lead to lower accuracy for internal depths that correspond to the structures of weak color contrast. To alleviate this issue, one may consider to replace the computation of intrinsic images [14] with an advanced CNN-based segmentation technique (e.g. [76]).

In the future, we will exploit our automatic depth map synthesis in image quality assessment task, where the depth information improved the state-of-the-art significantly [77]. We believe that other fields such as image completion, in painting, restoration, and panorama stitching will benefit from automatically generated depth maps as well.

Acknowledgments

This work was supported by V3C – "Visual Computing Competence Center" by Technology Agency of the Czech Republic, project no. TE01020415; by the Ministry of Education, Youth and Sports of the Czech Republic from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center – LM2015070"; by the Fulbright Commission in the Czech Republic; and by the ITRC program (IITP-2018-2016-0-00312, MSIT, Korea). We thank Martin Šimonovský for weighted edge detector, Jakub Krbec for implementing the applications, and Filip Šroubek and Jan Brejcha for valuable comments and fruitful discussions.

Appendix A. Spatially-variant point spread function

To assess the spatially-variant defocus kernel size b, we proceed as follows. The important variables are illustrated in Fig. A1. First, the *circle of confusion* (c) is calculated using the "Zeiss formula" (modern standard) [78]: $c = d_S/1500$, where d_S is the sensor diagonal size in millimeters. Then, the *hyperfocal distance* (H) is calculated as follows: $H = f^2/(N \cdot c)$, where f is focal length, N denotes the f-number, and c is the circle of confusion limit. The depth-of-field (DOF) is then: $DOF = D_F - D_N$, where D_F and D_N is the near- and the far-limit, respectively (the nearest and farthest distances in a scene that appear acceptably sharp in an image). $D_N = H \cdot S/(H + S)$, and $D_F = H \cdot S/(H - S)$, where H is hyperfocal distance, and S is focus distance. Finally, the kernel diameter (b) at the given point is calculated as follows: $b = \frac{f \cdot m_s}{N} \frac{x_d}{s \pm x_d}$, where m_s is magnification of the object in focus, and x_d is the distance between the current point from the focus plane. More specifically, $m_s = f/(s - f)$, and $x_d = |x - s|$. The diameter b is finally converted to pixels: $b_{px} = d_i/d_s \cdot b$, where d_i is the image diagonal in pixels.

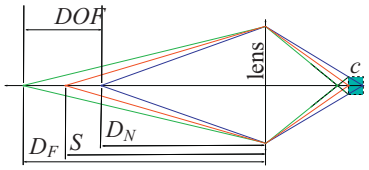

Fig. A1. Definition of variables and illustration of the depth-of-field for a symmetrical lens.

References

[1] Ng R, Levoy M, Brédif M, Duval G, Horowitz M, Hanrahan P. Light field photography with a hand-held plenoptic camera. Comput Sci Tech Rep CSTR 2005;2(11):1–11.
[2] Debevec PE, Malik J. Recovering high dynamic range radiance maps from photographs. In: Proceedings of the ACM SIGGRAPH classes. ACM; 2008. p. 31.
[3] Levin A, Fergus R, Durand F, Freeman WT. Image and depth from a conventional camera with a coded aperture. ACM Trans Gr 2007;26(3):70.
[4] Zheng Y, Chen X, Cheng M-M, Zhou K, Hu S-M, Mitra NJ. Interactive images: cuboid proxies for smart image manipulation. ACM Trans Gr 2012;31(4) 99:1–99:11.
[5] Chen T, Zhu Z, Shamir A, Hu S-M, Cohen-Or D. 3sweepp: Extracting editable objects from a single photo. ACM Trans Graph 2013;32(6) 195:1–195:10.
[6] Kholgade N, Simon T, Efros A, Sheikh Y. 3d object manipulation in a single photograph using stock 3d models. ACM Trans Graph 2014;33(4) 127:1–127:12.
[7] Hennessey JW, Mitra NJ. An image degradation model for depth-augmented image editing. Comput Gr Forum 2015;34(5):191–9.
[8] Thomee B., Shamma D.A., Friedland G., Elizalde B., Ni K., Poland D., et al. The new data and new challenges in multimedia research. arXiv:150301817 2015.
[9] Nayar SK, Narasimhan SG. Vision in bad weather. In: Proceedings of the IEEE international conference computer vision; vol.2. IEEE; 1999. p. 820–7.
[10] Lindeberg T, Garding J. Shape from texture from a multi-scale perspective. In: Proceedings of the international conference computer vision; 1993. p. 683–91.
[11] Zhang Z. Microsoft kinect sensor and its effect. IEEE MultiMedia 2012;19(2):4–10.
[12] Kopf J, Neubert B, Chen B, Cohen M, Cohen-Or D, Deussen O, et al. Deep photo: model-based photograph enhancement and viewing. ACM Trans Gr 2008;27(5):1–10.
[13] Chen B, Cohen M, Ramos G, Drucker S, Ofek E, Nister D. Interactive techniques for registering images to digital terrain and building models. Technical Report; 2008. https://www.microsoft.com/en-us/research/wp-content/uploads/2008/08/tr-2008-115.pdf.
[14] Bi S, Han X, Yu Y. An L_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM Trans Gr 2015;34(4):78.
[15] Sýkora D, Dingliana J, Collins S. As-rigid-as-possible image registration for hand-drawn cartoon animations. In: Proceedings of the international symposium non-photorealistic animation and rendering; 2009. p. 25–33.
[16] Levin A, Lischinski D, Weiss Y. A closed-form solution to natural image matting. IEEE Trans Pattern Anal Mach Intel 2008;30(2):228–42.
[17] Glocker B, Komodakis N, Tziritas G, Navab N, Paragios N. Dense image registration through MRFs and efficient linear programming. Med Image Anal 2008;12(6):731–41.
[18] Schechner YY, Kiryati N. Depth from defocus vs. stereo: How different really are they? Int J Comput Vis 2000;39(2):141–62.
[19] Rajagopalan A, Chaudhuri S, Mudenagudi U. Depth estimation and image restoration using defocused stereo pairs. IEEE Trans Pattern Anal Mach Intell 2004;26(11):1521–5.
[20] Kubota A, Aizawa K. Reconstructing arbitrarily focused images from two differently focused images using linear filters. IEEE Trans Image Process 2005;14(11):1848–59.
[21] Hasinoff SW, Kutulakos KN. A layer-based restoration framework for variable-aperture photography. In: Proceedings of the IEEE international conference computer vision; 2007. p. 1–8.
[22] Yang J, Schonfeld D. Virtual focus and depth estimation from defocused video sequences. IEEE Trans Image Process 2010;19(3):668–79.

[23] Yu L-F, Yeung S-K, Tai Y-W, Lin S. Shading-based shape refinement of rgb-d images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR); 2013.
[24] Bando Y, Nishita T. Towards digital refocusing from a single photograph. In: Proceedings of the pacific graphics; 2007. p. 363–72.
[25] Yan C-Y, Tien M-C, Wu J-L. Interactive background blurring. In: Proceedings of the ACM international conference multimedia; 2009. p. 817–20. ISBN 978-1-60558-608-3.
[26] Zhang W, Cham WK. Single-image refocusing and defocusing. IEEE Trans Image Process 2012;21(2):873–82.
[27] Levoy M, Hanrahan P. Light field rendering. In: Proceedings of the ACM SIGGRAPH; 1996. p. 31–42. ISBN 0-89791-746-4.
[28] Bishop T, Favaro P. The light field camera: extended depth of field, aliasing, and superresolution. IEEE Trans Pattern Anal Mach Intell 2012;34(5):972–86.
[29] Wanner S, Goldluecke B. Spatial and Angular Variational Super-Resolution of 4D Light Fields. In: Proceedings of the conference on computer vision–ECCV 2012, Part V. Springer Berlin Heidelberg; 2012. p. 608–21. ISBN 978-3-642-33715-4.
[30] Fiss J, Curless B, Szeliski R. Refocusing plenoptic images using depth-adaptive splatting. In: Proceedings of the IEEE international conference computational photography; 2014. p. 1–9. ISBN 978-1-4799-5188-8.
[31] Pujades S, Devernay F, Goldluecke B. Bayesian view synthesis and image-based rendering principles. In: Proceedings of the IEEE conference computer vision and pattern recognition; 2014. p. 3906–13. doi:10.1109/CVPR.2014.499.
[32] Levin A, Fergus R, Durand F, Freeman WT. Image and depth from a conventional camera with a coded aperture. ACM Trans Gr 2007;26(3).
[33] Moreno-Noguer F, Belhumeur PN, Nayar SK. Active refocusing of images and videos. ACM Trans Graph 2007;26(3).
[34] Sýkora D, Sedláček D, Jinchao S, Dingliana J, Collins S. Adding depth to cartoons using sparse depth (in)equalities. Comput Gr Forum 2010;29(2):615–23.
[35] Wang O, Lang M, Frei M, Hornung A, Smolic A, Gross M. StereoBrush: Interactive 2d to 3d conversion using discontinuous warps. In: Proceedings of eurographics symposium on sketch-based interfaces and modeling; 2011. p. 47–54.
[36] Iizuka S, Endo Y, Kanamori Y, Mitani J, Fukui Y. Efficient depth propagation for constructing a layered depth image from a single image. Comput Graph Forum 2014;33(7):279–88.
[37] Liao J, Shen S, Eisemann E. Depth map design and depth-based effects with a single image. In: Proceedings of the graphics interface; 2017. p. 57–64.
[38] Wang C-P, Wilson K, Snavely N. Accurate georegistration of point clouds using geographic data. In: Proceedings of the international conference 3D vision. IEEE; 2013. p. 33–40.
[39] Cho PL. 3d organization of 2d urban imagery. In: Proceedings of the applied imagery pattern recognition workshop. IEEE; 2007. p. 3–8.
[40] Li Y, Snavely N, Huttenlocher DP, Fua P. Worldwide pose estimation using 3d point clouds. In: Proceedings of the large-scale visual geo-localization. Springer; 2016. p. 147–63.
[41] Baboud L, Čadík M, Eisemann E, Seidel H-P. Automatic photo-to-terrain alignment for the annotation of mountain pictures. In: Proceedings of the conference computer vision and pattern recognition; 2011. p. 41–8. ISBN 978-1-4577-0394-2. doi:10.1109/CVPR.2011.5995727.
[42] Bae S, Durand F. Defocus Magnification. Comput Gr Forum 2007;26(3):571–9.
[43] Zhou C, Nayar S. What are good apertures for defocus deblurring?. In: Proceedings of the international conference computational photography; 2009. p. 1–8.
[44] Cook RL, Porter T, Carpenter L. Distributed ray tracing. ACM Comput Gr 1984;18(3):137–45.
[45] Lee S, Eisemann E, Seidel H-P. Real-time lens blur effects and focus control. ACM Trans Gr 2010;29(4) 65:1–7.
[46] Tan RT. Visibility in bad weather from a single image. In: Proceedings of the IEEE conference computer vision and pattern recognition; 2008. p. 1–8. ISBN 978-1-4244-2242-5.
[47] Fattal R. Single image dehazing. ACM Trans Gr 2008;27(3) 72:1–72:9.
[48] He K, Sun J, Tang X. Single image haze removal using dark channel prior. IEEE Trans Pattern Anal Mach Intell 2011;33(12):2341–53.
[49] Nishino K, Kratz L, Lombardi S. Bayesian defogging. Int J Comput Vision 2012;98(3):263–78.
[50] Meng G, Wang Y, Duan J, Xiang S, Pan C. Efficient image dehazing with boundary constraint and contextual regularization. In: Proceedings of the IEEE international conference computer vision; 2013. p. 617–24. ISBN 978-1-4799-2840-8.
[51] Fattal R. Dehazing using color-lines. ACM Trans Gr 2014;34(1) 13:1–13:14.
[52] Li Z, Tan P, Tan RT, Zou D, Zhou SZ, Cheong LF. Simultaneous video defogging and stereo reconstruction. In: Proceedings of the conference computer vision and pattern recognition; 2015. p. 4988–97.
[53] Tang K, Yang J, Wang J. Investigating haze-relevant features in a learning framework for image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 2995–3002. ISBN 978-1-4799-5118-5.
[54] Zhu Q, Mai J, Shao L. A fast single image haze removal algorithm using color attenuation prior. IEEE Trans Image Process 2015;24(11):3522–33.
[55] Cai B, Xu X, Jia K, Qing C, Tao D. Dehazenet: an end-to-end system for single image haze removal. 2016; abs/1601.07661. CoRR, http://arxiv.org/abs/1601.07661.
[56] Baatz G, Saurer O, Köser K, Pollefeys M. Large scale visual geo-localization of images in mountainous terrain. In: Proceedings of the European conference computer vision. Springer; 2012. p. 517–30.
[57] Lin TY, Belongie S, Hays J. Cross-view image geolocalization. In: Proceedings of the IEEE conference computer vision and pattern recognition; 2013. p. 891–8. ISBN 978-0-7695-4989-7. doi:10.1109/CVPR.2013.120.
[58] Hays J, Efros AA. IM2GPS: Estimating geographic information from a single image. In: Proceedings of the IEEE conference computer vision and pattern recognition; 2008. ISBN 9781424422432.
[59] Weyand T, Kostrikov I, Philbin J. PlaNet - Photo geolocation with convolutional neural networks 2016; 1602.05314: http://arxiv.org/abs/1602.05314.
[60] Arandjelović R, Gronat P, Torii A, Pajdla T, Sivic J. NetVLAD: CNN architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Washington, D.C, USA: IEEE Computer Society Press; 2016. p. 5297–307. ISBN 978-1-4673-8851-1.
[61] Brejcha J, Čadík M. State-of-the-art in visual geo-localization. Pattern Anal Appl 2017:1–25. doi:10.1007/s10044-017-0611-1.
[62] Dollar P, Zitnick CL. Structured forests for fast edge detection. In: Proceedings of the international conference computer vision; 2013. p. 1841–8. ISBN 978-1-4799-2840-8. arXiv: 1406.5549v1.
[63] Godard C, Mac Aodha O, Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the CVPR; 2017.
[64] Revaud J, Weinzaepfel P, Harchaoui Z, Schmid C. Deepmatching: hierarchical deformable dense matching. Int J Comput Vis 2016;120(3):300–23.
[65] Pan J, Hu Z, Su Z, Yang M-H. Deblurring text images via L0-regularized intensity and gradient prior. IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH; 2014. p. 2901–8. doi:101109/CVPR2014371.
[66] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. In: Proceedings of the advances in neural information processing systems. Curran Associates, Inc; 2014. p. 2366–74.
[67] Montgomery DC, Runger GC. Applied statistics and probability for engineers. John Wiley and Sons; 2003.
[68] Pertuz S, Puig D, Garcia MA. Analysis of focus measure operators for shape-from-focus. Pattern Recognit 2013;46(5):1415–32.
[69] Joshi N, Szeliski R, Kriegman DJ. PSF estimation using sharp edge prediction. In: Proceedings of the IEEE conference computer vision and pattern recognition; 2008. p. 1–8.
[70] Fergus R, Singh B, Hertzmann A, Roweis ST, Freeman WT. Removing camera shake from a single photograph. ACM Trans Gr 2006;25(3):787–94.
[71] Joshi N, Zitnick CL, Szeliski R, Kriegman DJ. Image deblurring and denoising using color priors. In: Proceedings of the IEEE conference computer vision and pattern recognition; 2009. p. 1550–7. ISBN 978-1-4244-3992-8.
[72] Shan Q, Jia J, Agarwala A. High-quality motion deblurring from a single image. ACM Trans Gr 2008;27(3) 73:1–73:10.
[73] Sorel M, Sroubek F, Flusser J. Super-resolution imaging. Towards Super-Resolution in the Presence of Spatially Varying Blur Digital Imaging and Computer Vision. CRC Press; 2010. p. 187–217. ISBN 978-1-4398-1930-2.
[74] Hertzmann A, Jacobs CE, Oliver N, Curless B, Salesin DH. Image analogies. In: Proceedings of the ACM SIGGRAPH; 2001. p. 327–40. ISBN 1-58113-374-X.
[75] Fišer J, Jamriška O, Lukáč M, Shechtman E, Asente P, Lu J, et al. StyLit: Illumination-Guided Example-Based Stylization of 3D Renderings. ACM Trans Graph 2016;35(4).
[76] He K, Gkioxari G, Dollár P, Girshick RB. Mask R-CNN. In: Proceedings of the IEEE international conference computer vision; 2017. p. 2980–8.
[77] Herzog R, Čadík M, Aydın TO, Kim KI, Myszkowski K, Seidel H-P. NoRM: no-reference image quality metric for realistic image synthesis. Comput Gr Forum 2012;31(2):545–54.
[78] Fleischer Kornelius J. Depth of Field An Insider's Look. Camera Lens News #1. Oberkochen, Germany: Carl Zeiss AG., Camera Lens Division; 1997.