Evaluating the Impact of Wide-Angle Lens Distortion on Learning-based Depth Estimation

Julie Buquet*[1 2], Jinsong Zhang[1] , Patrice Roulet[2], Simon Thibault[1], Jean-François Lalonde[1]
[1] Université Laval, Québec, QC, Canada
[2] Immervision, Montréal, QC, Canada

Abstract

Most computer vision research focuses on narrow angle lenses and is not adapted to super-wide-angle (aka spherical) lenses. This is mainly because current neural networks are not designed or trained to interpret the significant barrel distortion that is introduced in the captured image by such wide angle lenses.As these lenses capture a half-sphere or a section of sphere on the object space, barrel distortion appears when the image is projected on a 2D flat image sensor. By controlling this distortion at the lens design stage, camera designers can create some areas with augmented resolution [26]. In this work, we present an analysis of the impact of such augmented resolution on computer vision algorithm accuracy, using the problem of single image depth estimation as a case study. To this end, 360° *panorama datasets are warped to simulate different wide-angle lens datasets, which are then used to train identical neural networks. Each lens presents specific areas of the image with augmented resolution using spatially-varying non-linear distortion. We show that this property leads to better local accuracy in depth estimation. We also demonstrate that considering lens manufacturing improves performance when tested on realistic lenses, especially in the area of augmented resolution. We further show that this property helps to locally come closer to performances obtained on perspective images without cropping the field of view.*

1. Introduction

In the field of computer vision, learning-based approaches, especially those based on deep learning, have proved to be very efficient for many computer vision tasks. Indeed, trained Convolutional Neural Networks (CNNs) can now estimate depth [19, 9, 3, 7, 1], segment objects [11, 4] , and even infer 3D content from as little as a single image [18]. In turn, this has had impact on a variety of fields such as autonomous driving, medical operation assistance, consumer electronics and surveillance.

Given the ubiquity of "narrow angle" lenses (lenses that are well-approximated by the pinhole projection model), the vast majority of the work done in this area has focused on images with a narrow field of view and little or no lens distortion. However, the recent democratization of wide-angle imaging systems has allowed practitioners to easily capture images with much larger fields of view. Thanks to a highly negative meniscus as first element of the system, it is possible to render panoramic images higher than 90° (up to 360° for some systems). While this is beneficial for computer vision since a larger field of view should provide more complete information for improved scene understanding, these lenses also create distortion, that will modify the magnification and vary the pixel density across the image. Irrespective of whether this pixel density is linear or non-linear with respect to the field of view, it will cause straight lines to appear curved and objects to be warped. In turn, this creates a significant drop in performance when applying deep learning algorithms trained on perspective images since they have never seen such distortions during training. One possible solution is to train on lens-specific profiles, but this is prohibitive due to the challenge of acquiring sufficient training data. This problem is exacerbated by the fact that each lens model has a different distortion profile. Some works focus on adapting the network architecture and especially transforming the convolutional filters to apply them directly on spherical images without creating the distortion caused by the 2D flat projection [32]. Another solution involves removing the distortion via an image rectification (dewarping) process. This step often involves a 2D projection to recover the corresponding perspective image. However, this leads to pixel stretching and a field of view crop on the periphery (corners) of the image, resulting in a consequent loss of information compared to the original wide-angle image.

We are therefore left with the question: what is the impact of lens distortion profiles on computer vision algorithm accuracy? In this paper, we provide an answer to this question by using the problem of depth estimation from a single

*Corresponding author: jubuq@ulaval.ca

image [9, 3, 7] as a case study. In particular, we present two experiments that focus on spatially-varying, non-linear distortion profiles that augment the resolution locally. First, we aim at determining whether this increase in resolution does correlate with better depth estimation performance. Second, the same lens model might have variations in wide angle lens profiles, due to manufacturing tolerance. We will explore whether manufacturing tolerance has an impact on depth estimation. In both cases, our experiments reveal that lens distortion has a significant impact on performance, and that taking this into account during training is critical. With these observations, we finally estimate the drop in performances when we compare such networks with networks trained on perspective images. We compare the results between non-linear distortion profile in reference to the perspective images. In this last experiment we try to determine the impact of the dewarping process on depth estimation accuracy.

2. Related works

2.1. Single image depth estimation

Many techniques have been proposed for estimating depth from a single image over the past several years. The task is to predict, for each pixel in the input RGB image, the corresponding depth of the scene (either relative or absolute) at that pixel. Originally using graph-based algorithms such as MRFs [8, 21], recent approaches rely on deep learning. Approaches differ by the cues used to determine loss functions. These range from using stereo camera pairs [5, 30] to exploiting camera motion through the scene [34, 16, 6], or exploiting different types of data such as computer-generated images [23] or YouTube videos [15]. In this work, we exploit the recently proposed depth estimation algorithm of Hu et al. [9], which offers good performance when dataset providing ground truth per-pixel depth is available.

2.2. Application to wide angle

Wide angle imaging systems render large scenes in a single spherical image with a field of view higher than 80° making them useful for scene understanding. However, most of the methods noted above focus only on narrow angle images following the perspective projection (aka perspective images). Directly applying image rectification with 2D projections on wide-angle images would result in a loss of field of view or a lack of accuracy with the appearance of barrel distortion or discontinuities. Because of the high resolution of these systems, directly training networks on spherical images is also an issue because it requires annotated data with such resolution. Recently, 3D contents has been democratized and annotated data is increasingly available. Some works [36, 35] created spherical datasets using these 3D annotated data, making them able to train new networks. Other works [24] have developed networks able to transfer kernels from CNNs trained with perspective images to equirectangular panoramic images [31] or networks able to take two different projections of 360° images [28]. Moreover, some works focused on extracting 360° image features to produce the same output as a flat convolutional filter without 2D reprojection needed [32].

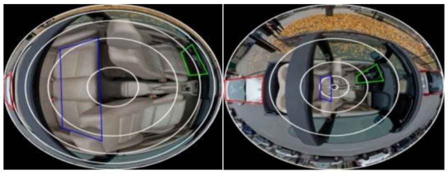


Figure 1. Images generated with two different panomorph lenses. Left: resolution is higher at the center, so the car interior (blue square) appears larger. Right: the augmented resolution on the edges makes the detection of the car body (red square) easier.

2.3. Controlled non-linear distortion

One of the main issues on wide-angle imaging is to enhance lens design parameters [20, 33], and/or use software processing [22] to reduce distortion. Panomorph lenses optimize the distortion using three important properties : anamorphose, freeform and controlled non-linear distortion [14]. Freeform and anamorphose are the ability for a lens to go beyond traditional rotationally symmetric image, for example having different horizontal and vertical field of view or different transverse and lateral magnification. Thanks to this property it is possible to optimize pixel density, sensor coverage and have areas of the image with higher pixel density meaning better resolution to define an object in this area. For example, Fig. 1 presents a freeform super-wide-angle lens producing an elliptic shape that covers more of the sensor than would a circular representation (obtained with a traditional rotationally symmetric wide-angle lens). In addition, by controlling the non-linear distortion function, it is possible to design a lens choosing the area of augmented resolution as explained in [26]. Such effects are visible on Fig. 1 where the car body (red square) is highly compressed on the lens with augmented resolution towards the center, while it is visible on the right where the resolution towards the edges has been augmented. As different applications have different regions of interest, this property leads to improved and customized imaging in different fields and for different applications, among them surveillance [25], consumer electronics [27] and medical assistance [17].

3. Methodology

As some characteristics of the image can affect depth estimation, we studied the influence of spatially-varying non-

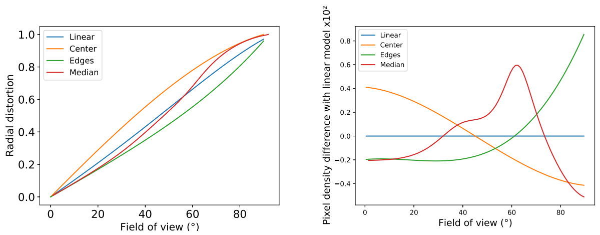


Figure 2. Left : Distortion functions of linear (blue), center- (orange), edges- (green), and median-augmented (red) lenses used in this evaluation. The radial distortion r corresponds to the relative distance from the center of the image x_0. Each pixel on the same circle of radius $r < 1$ with center x_0 correspond to the same field of view in the scene. One should note that the pixel density follows the same tendency meaning it is more important on the edges of the image for all lens models. Right : Pixel density difference with the linear model (blue) of the center- (orange), edges- (green) and median-augmented (red) lenses used. We computed the derivative of the radial distortion for each lens model and then subtracted the result for the linear model. Sections of curves that are above the blue line have higher pixel density than the linear model at this field of view (e.g. the orange curve for the center-augmented model is above 0 for lower value of field of view).

linear distortion of panomorph lenses on depth estimation. We first compared them to a linear fisheye lens imaging the same 180° field of view and then to their corresponding 140° perspective projection with distortion correction (image rectification). Deep CNNs trained to estimate depth from a single image were used.

3.1. Lenses

Four different lenses are studied. First, a fisheye lens with a linear distortion profile is used as a reference for the first experiment. Second, three panomorph lenses are used: the first has augmented resolution on the center of the image, the second on the edges and the last on the median area. Their distortion functions are presented in Fig. 2 (left). The curves also show that for each lens model, the higher pixel density is still located on the edges of the image[1]. Fig. 2 (right) shows the difference of pixel density with the linear model. The center-augmented model curve (orange) is above 0 (blue line) for lower field of view, which means the pixel density is increased in this area.

3.2. Dataset generation

Working on such images makes it hard to have available data for depth estimation. Instead, we simulated each of these lens designs by warping 360° panoramic images from two existing datasets: SunCG [23] and Matterport3D [2]. Both contain 360° photographs of interior scenes along with the corresponding ground truth depth estimation. SunCG contains computer-generated images, while Matterport3D captures real images.

The panoramic RGB image and its corresponding depth map are warped to wide angle images according to different distortion profiles [29]. For this, a mesh is first built on the desired wide-angle image with coordinates (u, v). Each such coordinate is then backprojected according to the lens distortion profile to obtain a 3D point (x, y, z), which is subsequently re-projected into a second mesh (u', v') on the panoramic image. The color/depth value for (u, v) is obtained via bilinear interpolation. Fig. 3 shows example images obtained with this technique, showing that different lens profiles can accurately be reproduced from the same 360° panoramic input. After warping, some pixels are not exploitable, for SunCG depth is not available for doors, windows, mirrors and are taken as infinitely far. The Matterport3D dataset does not have depth information for very dark regions, such as windows, mirrors or light sources, due to limitation of the sensor in the capture system.

Three different datasets (one for each lens profile) are thus computed. They are composed of 5585/675/675 images from SunCG and 8591/1137/1072 from Matterport3D for the training/validation/test datasets respectively. After dark pixel checking, Matterport3D remains unchanged while SunCG becomes 164/112/13. The validation set is a separated sample of data used during network training. Early stopping is used to terminate the training if the performance is saturated.

3.3. Network architecture

We borrow the architecture from Hu et al. [9], which is illustrated in Fig. 4 and briefly reviewed here for completeness. The network is built upon an encoder-decoder structure and aims at generating higher resolution depth map from a single RGB image. The architecture consists of four modules, including a SqueezeNet [10], AlexNet [13] and two DeConv (upsampling-convolution) blocks. The encoder network extracts features from a single RGB at different scales. We use SqueezeNet module to build the encoder network [10], which is known for its compact architecture and low memory usage. The decoder network takes the features with the smallest scale from the encoder network as input. This decoder network is built with deconvolution layers (upsampling and convolution), which outputs a global feature map that has the same spatial resolution as the image. The feature maps at different scales are also transmitted to another decoder network which aims at extracting local information by upsampling and merging all these feature maps into a local feature map. Both global and local feature maps are used to estimate the depth map via the AlexNet [13] module that will output a depth map. This network is composed of 196 layers and requires 596.45 MB

[1] These lenses are mostly used in surveillance applications; edges-augmented lenses are typically fixed on a wall while the others are often mounted on the ceiling.

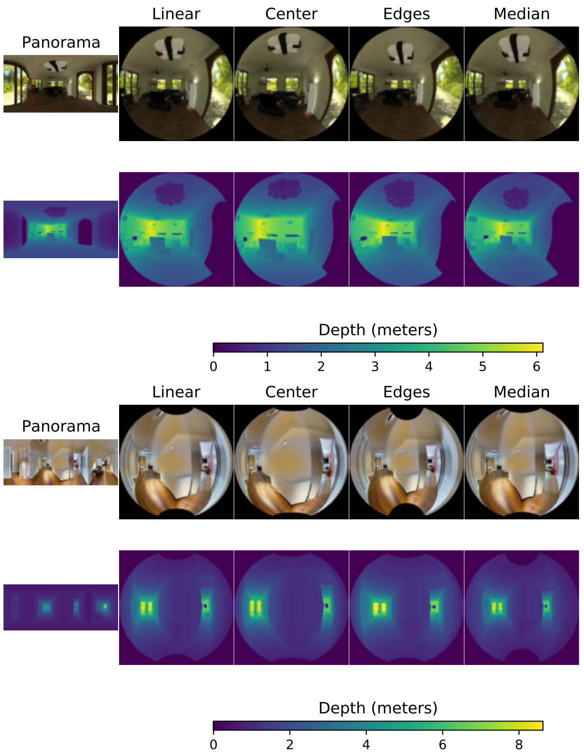


Figure 3. Examples of images from our datasets. From a 360° panoramic input (left, top) with the corresponding depth map (left, bottom), panoramic images with different distortion functions are generated by sampling the original panorama (from left to right: panorama, linear distortion, center-, edges- and median-augmented resolution). This study exploits data from two sources: SunCG [23] (top, synthetic) and Matterport [2] (bottom, real). For each dataset, both RGB images and depth are generated.

of memory usage.

3.4. Training procedure

To train the network, we use the combination of three loss functions for depth estimation proposed by Hu et al. [9]. The first one, $\mathcal{L}_{\text{depth}}$, directly estimates the log difference between the ground truth depth d_i at pixel i and the estimate $\tilde{d}_i$:

$$\mathcal{L}_{\text{depth}} = \frac{1}{N}\sum_{i=1}^{N} \ln(|d_i - \tilde{d}_i| + \alpha) \tag{1}$$

with $\alpha = 10^{-5}$ and N the number of pixels on the image. The log is here to give more importance to a pixel closer to a the camera for the same value of error. Two additional loss functions $\mathcal{L}_{\text{grad}}$ and $\mathcal{L}_{\text{normal}}$ are introduced to penalize the error around edges in the depth map:

$$\mathcal{L}_{\text{grad}} = \frac{1}{N}\sum_{i=1}^{N} \ln(\nabla_x(|d_i - \tilde{d}_i|) + \alpha) + \ln(\nabla_y(|d_i - \tilde{d}_i|) + \alpha), \tag{2}$$

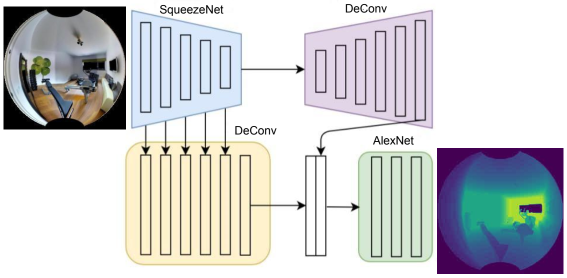


Figure 4. Structure of the depth estimation network used in this study, borrowed from Hu et al. [9]. The input is a single wide angle RGB image (here a linear fisheye from Matterport3D). The encoder (blue, which uses a SqueezeNet architecture as a backbone network [10]) extracts multi-scale features from the input image. The deconvolution layers (purple) takes the last scale features as input and aims to find global depth information. Features extracted at different scales from the encoder are combined through a series of deconvolution layers (yellow), which servers as local depth information. The AlexNet [13] takes both global and local features to estimate the depth map (bottom right).

$$\mathcal{L}_{\text{normal}} = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{\langle \mathbf{n}_i^d, \tilde{\mathbf{n}}_i^d \rangle}{\sqrt{\langle \mathbf{n}_i^d, \mathbf{n}_i^d \rangle}\sqrt{\langle \tilde{\mathbf{n}}_i^d, \tilde{\mathbf{n}}_i^d \rangle}}\,, \tag{3}$$

$$\text{with } \mathbf{n}_i = [-\nabla_x(d_i), -\nabla_y(d_i), 1]\,.$$

Where the $\nabla_*(d_i)$ indicates the spatial gradient at the i^{th} pixel along the x, y direction in the image plane. Both gradient $\mathcal{L}_{\text{gradient}}$ and normal $\mathcal{L}_{\text{normal}}$ losses are used, because they are sensitive to the error at different scales. The discontinuous boundary structure of objects is captured by the loss $\mathcal{L}_{\text{gradient}}$, while fine structures of a continuous surface are modelled by the loss $\mathcal{L}_{\text{normal}}$. We train the network using the Adam optimizer [12] with an initial learning rate of 0.0001 and a batch size of 32. In contrast to the vanilla Stochastic Gradient Descent (SGD) optimizer, the Adam optimizer is less sensitive to the choice of hyper-parameters and is widely used in learning.

Training the network for 650 epochs on a Titan X GPU takes approximately five days using samples with a resolution of 256×256 pixels.

4. Experiments

To estimate the influence of locally augmented resolution on depth estimation, we first compare the linear fisheye lens with two panomorph lenses : those with center- and edges- augmented resolution. Then, the effect of tolerancing in the lens profile on the CNN's robustness is studied on the lens profile with augmented resolution on the median zone of the image (see Sec. 3.1). A different network is trained for each of the four lens profiles (Sec. 3.1).In the

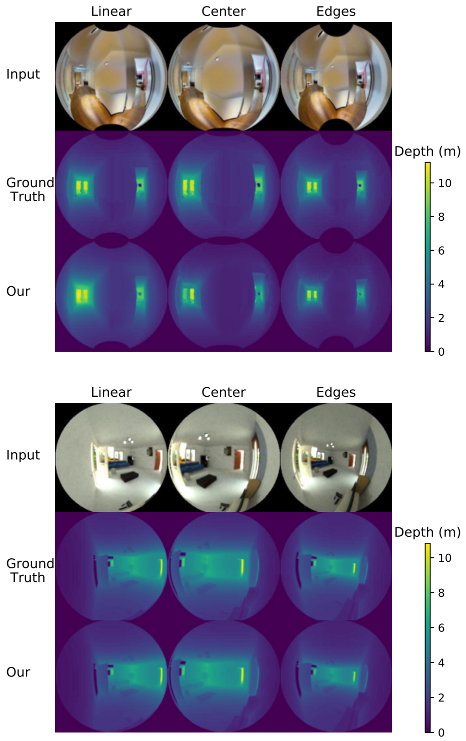


Figure 5. Qualitative depth estimation results on wide angle images generated from Matterport3D (top) and SunCG (bottom) . The first row is the input for each network. From left to right: linear distortion, augmented resolution on the center, augmented resolution at the edges. The second row is the ground truth and the last is the depth predicted by the network.

final experiment we compare the accuracy of a panomorph-trained network (with a lens profile with augmented resolution on the median zone) and a perspective-trained network. We want to estimate the loss in accuracy when we do not remove the distortion and keep the whole field of view on the input image.

4.1. Influence of the locally-augmented resolution

In this first experiment, 3 lens profiles and their networks are considered : the linear fisheye, the center- and edges-augmented lenses. During training, all the networks learned the same way whether the distortion was linear or not. Then the goal will be to study if we can notice the spatially varying distortion for each network at inference time.

Outputs from inference time, are presented in Fig. 5 and evaluated using two metrics: the RMSE and the relative error on each pixel of the image with $\tilde{d}_i$ corresponding to the depth on the output for the i^{th} pixel and d_i the corresponding ground truth.

Lens model	Linear	Center	Edges
RMSE	0.39	0.39	0.38
REL	0.24	0.29	0.22

Table 1. RMSE (Eq. 4) and REL (Eq. 5) metrics calculated on depth values for each lens profile. Metrics are reported in meters.

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{i=1}^{T}(d_i - \tilde{d}_i)^2} \tag{4}$$

$$\text{REL} = \frac{1}{T}\sum_{i=1}^{T}\frac{||d_i - \tilde{d}_i||}{\tilde{d}_i} \tag{5}$$

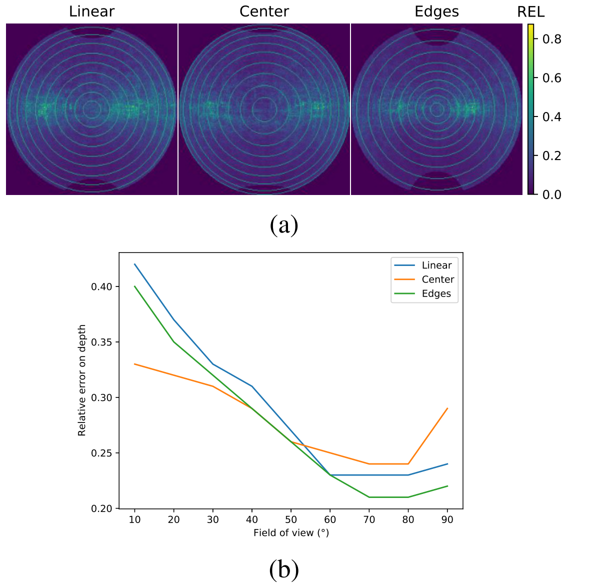
The average RMSE (Eq. 4) and REL (Eq. 5) metrics are calculated on the entire image (excluding the black regions in the corners and black pixels). Results presented in Tab. 1 provide global performance on the entire image. As we can see, they globally perform the same for both metrics and the non-linear distortion does not seem to significantly affect the accuracy when looking at the image in its entirety.

As the networks are globally comparable (Tab. 1), we then generated the average relative-error map for each network (Fig. 6). The network trained with centered-augmented resolution is more accurate at the center of the image than the other networks while the network with augmented resolution on the edges performs better for highest field of view values. In this way, the augmented resolution seems to locally influence the accuracy of the network so depending on the application, it is possible to have a better depth estimation on the area of interest by choosing the appropriate distortion function.

4.2. Influence of the tolerancing

So far, nominal distortion functions were used for each lens profile. In practice, however, errors in the lens manufacturing process can create deviations from these nominal curves. This second experiment studies the impact of tolerance on depth estimation.

For this experiment, we considered another panomorph lens with augmented resolution in the median zone presented in Sec. 3.1. From its nominal distortion function, we generated twenty other distortion functions (toleranced functions) according to tolerance specifications. The bias introduced in the distortion function is calculated from the tilt. In order to generate a bias, we randomly select a value for this tilt within a certain range given by tolerances, where this range depends on the dimension of the lens. Tolerancing ranges were chosen to fit those of commercialized wide angle lenses. Fig. 7 presents the bias introduced for 10 of those generated functions. The bias here is the difference on radial distortion to nominal distortion function of Fig. 2 (left).



(a)

(b)

Figure 6. Relative error (REL metric, Eq. 5) for each lens profile (lower is better). In (a), the mean relative error maps (in meters) computed over the entire test set are illustrated for each lens profile (left to right: linear distortion, augmented resolution on the center, augmented resolution on the edges). Each circle represents 10° of field of view. In (b), the REL is averaged over each circle of relative error calculated for each circle of 10°, and displayed on the same plot for better comparison. For low fields of view, the network trained with augmented resolution on the center (orange) performs better than the others while the network with augmented resolution on the edges (green) is the best for higher fields of view.

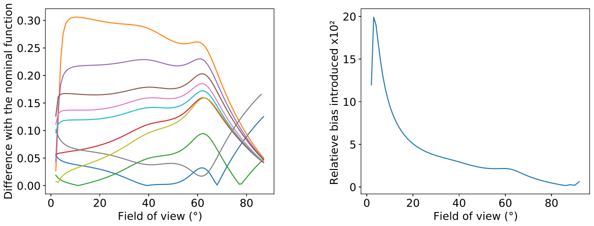


Figure 7. Generation of toleranced distortion functions and toleranced test dataset. Left :Example of bias introduced in the distortion functions generated. Each coloured curve represents the difference on radial distortion to the nominal function (Fig. 2) introduced for one generated function. Right : relative difference on radial distortion introduced in the test dataset compared to the nominal function. For each value of field of view the relative difference to the nominal function for all samples is calculated. The mean is printed

The first network was trained for 1200 epochs in the same way as before (Sec 3.4) using the nominal lens profile. A second network was trained considering tolerancing on the lens profile. During training time, this takes a panoramic RGB image as input and randomly choses one of the toleranced distortion function to warp the input. Each network took around 8 days to train.

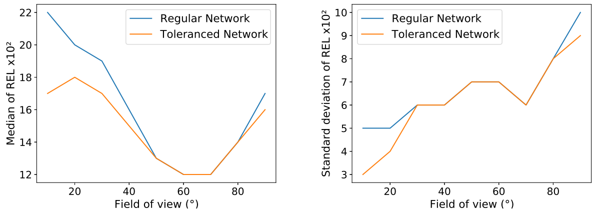


Figure 8. Analysis of regular (blue) and toleranced (orange) networks on a test dataset with tolerancing : median (left) and standard deviation (right) of the REL metric (Eq. 5) for each network. We observe that the network trained with tolerancing (blue) is better than the regular network (orange) especially where tolerancing introduced more bias.

At inference time, the panoramic test dataset of Sec. 3.2 was used. For each sample one of the toleranced functions is chosen to warp the image. Unlike tolerancing during training, here the pair image/distortion function is unique to create a fixed toleranced test dataset used on both networks. The bias presented in Fig. 7(right) represents the relative difference between the radial distortion of the nominal function (Fig. 2 in Sec. 3.1) and the average radial distortion from all samples of the toleranced test dataset for each field of view value. We locally observed the behaviour of each network to see if the regular network could still estimate depth correctly on such dataset and how any bias introduced in the test dataset influenced the accuracy of each network.

As shown in Fig. 8, the toleranced network globally performs better than the regular network because it was trained with toleranced datas. This is especially noticeable where more bias was introduced relative to the original radial distortion. Moreover, they are essentially equals in the area of augmented resolution around 60°.

4.3. Comparison with perspective images

In this section, we investigate the impact of the dewarping process on depth estimation accuracy. We compared networks trained with median-augmented panomorph images with identical networks trained with the corresponding perspective projection. Once again the training conditions were the same as in Sec. 3.4. Both networks were trained for 1200 epochs. The goal is to determine if the non-linear distortion can be used to locally reach the same accuracy as for perspective images while keeping the entire 180° field of view.

A “perspective” dataset was created by dewarping the wide-angle images (median-augmented) while limiting our field of view to 140°. Input images and their corresponding estimated depth maps are shown in Fig. 9. In this ex-

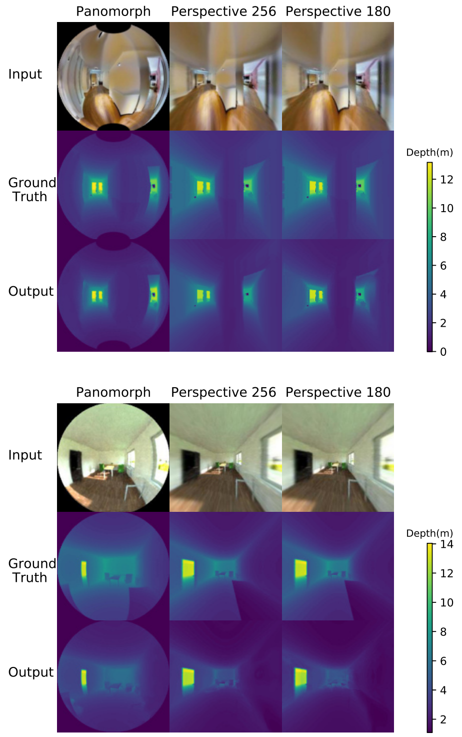


Figure 9. Qualitative depth estimation results on wide angle images and 140° perspective generated from Matterport3D (top) and SunCG (bottom). The first row is the input for each network. From left to right: Panomorph with augmented resolution on the median zone, Perspective projection with 140° (256 × 256), Perspective projection with 140° (180 × 180)

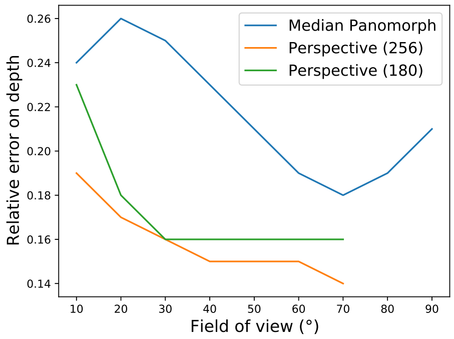


Figure 10. Relative error (REL metric, Eq. 5) for Panomorph with augmented resolution on the median zone, Perspective projection with 140° (256 × 256), Perspective projection with 140° (180 × 180). Networks trained on perspective images (orange and green) performs better than the one trained on Panomorph images (blue) regardless the resolution used. However the gap is reduced in the area of augmented resolution.

periment, wide-angle images were taken with the resolution of 256 × 256 pixels and two datasets of corresponding perspective images were created with different resolution : 256 × 256 and 180 × 180 pixels.This last resolution was chosen to have the same number of pixels for the overall 140° as in the wide angle image. We compared the relative error on depth estimation on the common field of view (0 to 140°) (Fig. 10). Even if the networks trained on perspective images (orange and green) globally perform better, the network trained on median-augmented wide angle images shows closer performances in his area of increased resolution (between 60° and 75° of half field of view). Looking at the perspective networks, it seems that the overall resolution does not affect consequently networks accuracy (orange and green curve are close to each other). However, the localized pixel density, which is higher at the edges of a perspective image (as the height on the image is a tangential function of the field of view for perspective images), seems to dramatically increase the accuracy. In the end, for each image type, the accuracy of depth estimation remains higher where the local pixel density is the highest, thereby, explaining how non-linear distortion leads to better localized depth estimation. We are currently working on determining if some areas of the wide-angle image could have a bigger pixel density compared to its perspective equivalent . It would be interesting to compare networks performance in such area to determine if the local resolution is the only factor impacting accuracy on depth estimation.

5. Conclusion

We built datasets of wide-angle images simulating spatially-varying non-linear distortion profiles, and we used these datasets to train CNNs for depth estimation. Since the lenses we used present different local areas of augmented resolution, this allowed us to determine that augmented resolution indeed corresponds to improved performance in depth estimation. In this way, controlling spatially-varying non-linear distortion leads to locally improved depth estimation. This property can improve network accuracy in the area of interest without losing accuracy elsewhere in the image or cropping a portion of the field of view. We also built a dataset with variations in lens profiles to take account of manufacturing tolerancing and saw that our network was still accurate. However, when training another network on data that included tolerancing information, the performance was further improved, especially in the areas where tolerancing induced higher variations. Finally we compared performance between networks trained and tested on panomorph images and perspective images. Even if the network trained on perspective images remains more accurate regardless the overall resolution, non-linear distortion on wide angle images seems to be helpful to get locally closer to such performances using panomorph images. In the end, we saw that for all datasets, local pixel density impacts depth estimation accuracy making non-linear distortion a powerful tool to locally enhance network performance. For a given application, it would be possible to estimate a distortion function that would help reach the desired accuracy on depth estimation without cropping the field of view. This could constitute a specification for lens designers to conceive cameras optimized for specific scene understanding tasks.

The main limitation of our work is that the analysis was performed entirely on simulated data, with a resolution limited to 256×256 pixels. We are currently working on collecting our own annotated images using a panomorph camera and structure from motion algorithm to constitute sparse ground truth depth maps. We could also investigate the impact of such distortion profiles on other network architectures for depth estimation or other applications such as object classification to enlarge the scope of applications. It would also be interesting to investigate different type of dewarping methods. As perspective projection presents drawbacks such as pixel stretching on the edges, some methods try to combine it with other type of projections in order to enlarge the field of view on the dewarped image, which could potentially alleviate this problem.

Acknowledgments

This work is supported by the Natural Sciences and Engineering Council of Canada (NSERC) Industrial Research Chair in Optical Design (IRCPJ 372500) and by Immervision who gave access to the panomorph lens profiles.

References

[1] A. Atapour-Abarghouei and T.P. Breckon. Real-time monocular depth estimation using synthetic datawith domain adaptation via image style transfer. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Nießner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*, 2017.

[3] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[4] R. Girshick. Fast r-cnn. In *Int. Conf. Comput. Vis.*, 2015.

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[6] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Int. Conf. Comput. Vis.*, 2019.

[7] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Int. Conf. Comput. Vis.*, 2019.

[8] Derek W Hoiem, Alexei A. Efros, and Martial H Hebert. Automatic photo pop-up. In *SIGGRAPH*, 2005.

[9] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE, 2019.

[10] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[11] P. Dollar K. He, G. Gkioxari and R. Girshick. Mask r-cnn. In *Int. Conf. Comput. Vis.*, 2017.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutionalneural networks. In *Adv. Neural Inform. Process. Syst.*, 2012.

[14] Bastien Martin Larivière. *Amélioration des performances des systèmes d'imagerie panoramiques et Panomorphes*. PhD thesis, Université Laval, 2014.

[15] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[16] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018.

[17] P. Roulet, P. Konen, M. Villegas, S. Thibault, and Pierre Y. Garneau. 360° endoscopy using panomorph lens technology. *SPIE Proceedings : Endoscopic Microscopy V*, 7558, 2010.

[18] C.Rupprecht S. Wu and A. Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[19] G. Brazil S. Zhu and X. Liu. The edge of depth: Explicit constraints between segmentation and depth. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[20] F. E. Sahin and A. R. Tanguay Jr. Distortion optimization for wide-angle computational cameras. *Optics Express*, 26(5):5478–5487, March 2018.

[21] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Depth perception from a single still image. In *AAAI*, 2008.

[22] Yichang Shih, Wei-Sheng Lai, and Chia-Kai Liang. Distortion-free wide-angle portraits on camera phones. *ACM Transactions on Graphics (SIGGRAPH)*, 2019.

[23] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016.

[24] Yu-Chuan Su and Kristen Grauman. Kernel transformer networks for compact spherical convolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[25] S. Thibault. Enhanced surveillance system based on panomorph panoramic lenses. In *The International Society for Optical Engineering 6540*, volume 6540, 2007.

[26] S. Thibault, J. Gauvin, M. Doucet, and M. Wang. Enhanced optical design by distortion control. In *The International Society for Optical Engineering*, 2005.

[27] S. Thibault, J. Parent, H. Zhang, X. Du, and P. Roulet. Consumer electronic optics: How small a lens can be ? the case of panomorph lenses. In *Proc. SPIE 9192, Current Developments in Lens Design and Optical Engineering XV, 91920H (25 September 2014)*, 2014.

[28] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360○ depth estimation via bi-projection fusion. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[29] Jianxiong Xiao. 3d geometry for panorama.

[30] Junyuan Xie and Ross Girshick. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks. In *Eur. Conf. Comput. Vis.*, 2016.

[31] K. Grauman Y. Su. Kernel transformer networks for compact spherical convolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[32] K. Grauman Y. Su. Learning spherical convolution for fast features from 360° imagery. In *Adv. Neural Inform. Process. Syst.*, 2017.

[33] Miao Zhang, Yongri Piao, Nam-Woo Kim, and Eun-Soo Kim. Distortion-free wide-angle 3d imaging and visualization using off-axially distributed image sensing. *Optics letters*, 39(14):4212–4214, July 2014.

[34] Tinghui Zhou, Matthew Brown, Noah Snavely, and David Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[35] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360° depth estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[36] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Eur. Conf. Comput. Vis.*, 2018.