Supplementary Material: Convolution kernel adaptation to calibrated fisheye

Bruno Berenguel-Baeta*
berenguel@unizar.es
Maria Santos-Villafranca*
m.santos@unizar.es
Jesus Bermudez-Cameo
bermudez@unizar.es
Alejandro Perez-Yus
alperez@unizar.es
Jose J. Guerrero
josechu.guerrero@unizar.es

Instituto de Investigacion en Ingenieria de Aragon (I3A)
Universidad de Zaragoza
Zaragoza, Spain

1 Image rectification

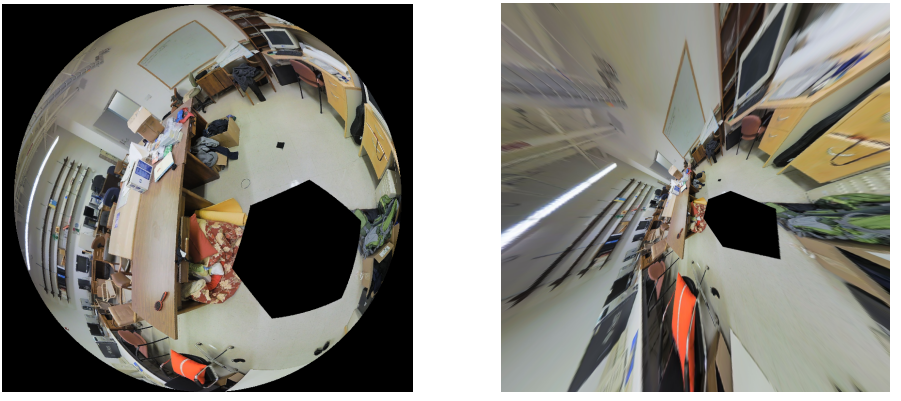
One of the main questions that may rise after reading our article is: *Why don't you rectify the fisheye image and use the Convolutional Neural Networks for perspective images on the rectified image?* The answer for this question comes in two parts:

1. Not all the wide-field-of-view images can be rectified. Cameras with a field of view equal or greater than 180° cannot be rectified completely, since the perspective projection model has a limit at this field of view. That means that we may loss part of the environment information in the rectification process.

2. Although the image rectification corrects the radial distortion, when having wide field of view the effect of projecting onto a plane results in large deformations in regions far from the principal point. In some cases, this projective deformation modifies the receptive field of the CNNs even more than the original radial distortion.

In the main article, we present results of two different fisheye camera calibrations, each corresponding to a different field of view. For one of the cameras, this rectification process is impossible to achieve completely (i.e. the camera has a field of view of 195°) while with the other, with a field of view of 165°, this rectification is possible but the rectified images are greatly deformed (see Fig. 1).

* Equal contribution



(a) Fisheye image (b) Rectified image

Figure 1: Comparison of a fisheye image from the dataset F165 and the rectification of the image in the perspective projection model.

	Input	MRE ↓	MAE ↓	RMSE ↓	RMSE_{log} ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
BL	*F165*	0.2670	0.3790	0.5005	0.2324	0.4377	0.7198	0.8579
	F165-R	0.8595	0.6412	0.9714	0.4178	0.3000	0.5509	0.7305
FT	*F165*	0.2508	0.3582	0.4040	0.1899	0.4962	0.7628	0.8879
	F165-R	0.7758	0.5999	0.8933	0.3661	0.3016	0.5710	0.7618

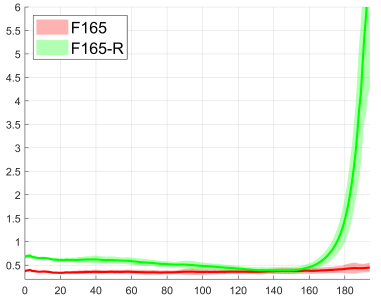
Table 1: Monocular depth estimation **with standard convolutions** for U-Net neural network with fisheye (F165) and rectified (F165-R) input images. BL: Base Line; FT: Fine Tuned.

We compare the performance of the CNN based on standard convolutions when directly using the fisheye image as input (F165), as considered in the main article, versus using a rectified version of the image as input (F165-R). The evaluation is made rectifying the whole input image, estimate depth or semantic segmentation and un-rectify the output of the network to compare with ground truth in the fisheye domain.

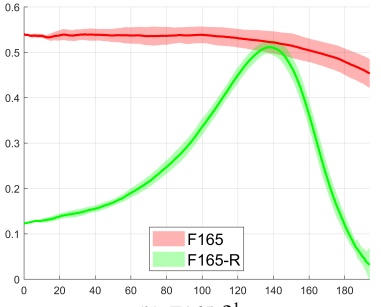
From the results of this evaluation, shown in Table 1 for depth estimation and Table 2 for semantic segmentation, we observe that rectify and un-rectify the fisheye images is not a good option for the proposed tasks with a CNN. The great deformation introduced in the rectification process seems to worsen the performance of the network with standard convolutions. This behaviour is also observable in Figure 2 and Figure 3, where metrics are presented with respect the radius of the fisheye image. In the depth estimation, the rectified image provides worse results in the outer part of the image. Besides, in the δ^1 metric, the behavior of the network presents better results in the middle range of the image radius, where the deformation in the rectified image is lower, and a really bad performance in the center and border of the image. In the semantic segmentation comparison, we observe a similar

	Input	mIoU	mAcc
BL	*F165*	15.12	24.36
	F165-R	11.81	19.82
FT	*F165*	27.70	36.51
	F165-R	21.99	29.61

Table 2: Semantic segmentation **with standard convolutions** for U-Net neural network with fisheye (F165) and rectified (F165-R) input images. BL: Base Line; FT: Fine Tuned.

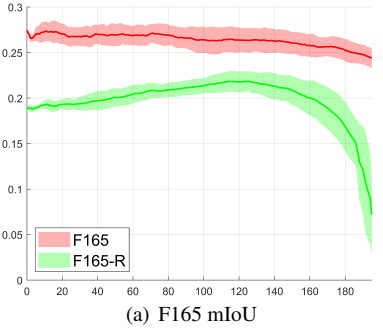


(a) F165 RMSE

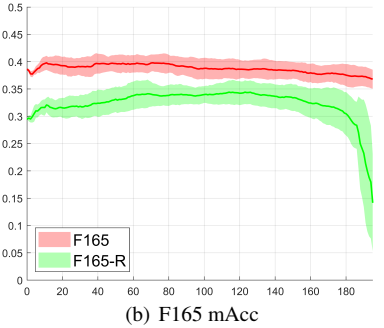


(b) F165 δ^1

Figure 2: Comparison and results of depth estimation with the U-Net like network **with standard convolutions** on the fisheye (F165) and rectified (F165-R) images. The x-axis defines the distance of the pixels to the optical center and the y-axis the computed error, defined as mean and one standard deviation.



(a) F165 mIoU



(b) F165 mAcc

Figure 3: Comparison and results of semantic segmentation with the U-Net like network **with standard convolutions** on the fisheye (F165) and rectified (F165-R) images. The x-axis defines the distance of the pixels to the optical center and the y-axis the computed error, defined as mean and one standard deviation.

behaviour for both approaches (directly use the fisheye and with rectification). While the use of standard convolutions in the fisheye image directly, as in the main paper, outperforms the rectification approach, both methods have worse performance at the border of the image. In this case, it is better to work directly on the fisheye image than rectifying it and use it in the network.