A comparative study of semantic segmentation using omnidirectional images

Ahmed Rida Sekkat, Yohan Dupuis, Paul Honeine, Pascal Vasseur

A comparative study of semantic segmentation using omnidirectional images

Ahmed Rida Sekkat[1], Yohan Dupuis[2], Paul Honeine[1] and Pascal Vasseur[1].

[1]Normandie Univ, UNIROUEN, LITIS, Rouen, France
[2]Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, Rouen, France

Abstract. The semantic segmentation of omnidirectional urban driving images is a research topic that has increasingly attracted the attention of researchers. This paper presents a thorough comparative study of different neural network models trained on four different representations: perspective, equirectangular, spherical and fisheye. We use in this study real perspective images, and synthetic perspective, fisheye and equirectangular images, as well as a test set of real fisheye images. We evaluate the performance of convolution on spherical images and perspective images. The conclusions obtained by analyzing the results of this study are multiple and help understanding how different networks learn to deal with omnidirectional distortions. Our main finding is that models trained on omnidirectional images are robust against modality changes and are able to learn a universal representation, giving good results in both perspective and omnidirectional images. The relevance of all results is examined with an analysis of quantitative measures.

Keywords: Omnidirectional, Equirectangular, Fisheye, Deep Convolutional Neural Networks, Semantic Segmentation

1 Introduction

Thanks to their large field-of-view, omnidirectional images are omnipresent in intelligent vehicles and robot navigation systems. At the same time, deep learning for computer vision tasks has never been used as much as it is currently. However, computer vision algorithms used in these systems and vehicles for tasks like scene understanding are mostly developed and tested for perspective conventional images. Hence the importance of optimizing these algorithms for omnidirectional imaging. We can notice a recent growing interest in this research subject. Several works treated the adaptation of existing algorithms or the development of new ones for tasks like object recognition and semantic segmentation on omnidirectional images, such as 360° and fisheye. In these two tasks, deep learning using convolutional neural networks (CNNs) on perspective images is the state-of-the-art solution. This is mainly thanks to the emergence of large-scale datasets of perspective images with ground truth annotation, such as CamVid [3] and Cityscapes [8]. This convenience is not available for omnidirectional images. Until now, there is no available dataset of omnidirectional

real urban driving images with ground truth. To compensate this major issue, several contributions on semantic segmentation of fisheye images work on data augmentation by training the state-of-the-art CNNs on perspective images that were deformed with a distortion simulating a fisheye effect [28, 11, 10]. On the other hand, some researchers proposed to encode directly the omnidirectional representation in the CNN [6]. More works proposed CNNs with deformable kernels [9, 16], or used icosahedron spherical image representation and spherical CNNs [12, 17].

More recently, researchers are considering the generation of synthetic images with realistic textures, thanks to simulators like CARLA simulator and Grand Theft Auto V (GTA V), which is a very high-quality video game. The recently published OmniScape Dataset [29] contains synthetic perspective, fisheye, catadioptric and 360° urban driving images with ground truth rendered from a virtual city and comes with pixel-level semantic annotation. In this work we take advantage of this dataset and CamVid, as well as a test set of real fisheye images that we captured and manually annotated, in order to make a study of some state-of-the-art semantic segmentation networks. This study consists in quantitative comparative analyses using semantic segmentation task to take stock of research progress and answer the following questions:

- Is training on omnidirectional representations sufficient to have good results? Or do we need to adapt CNNs for omnidirectional images?
- Do networks learn a universal representation when trained on omnidirectional images? And what are their performances on perspective images?
- Do spherical convolutions give better results than conventional convolutions?

In order to answer these questions, we conduct several experiments using a set of OmniScape synthetic images with perspective, fisheye and equirectangular projection of the same scene taken from both front sides of a motorcycle, images from CamVid dataset and our test set of real annotated fisheye images. First, we test several semantic segmentation networks on CamVid images and choose the four networks that give the best results. We then make a cross-modality experiment by retraining these networks separately on CamVid images, OmniScape perspective, fisheye, and equirectangular images, to test them one by one on all these representations, as well as on our test set of real fisheye images. We also use a convolutional network for spherical images to perform semantic segmentation using the same equirectangular images used in the previous experiments. At the end, this allows us to conclude on the efficiency of state-of-the-art neural networks dedicated to semantic segmentation of perspective images on equirectangular and fisheye images, as well as the performance of these networks when trained on omnidirectional images. And finally, the relevance of spherical CNNs is compared to these networks. Since dataset of omnidirectional urban driving images with ground truth were not available, studies made on semantic segmentation of real fisheye images rarely present quantitative measures, hence the interest to make quantitative evaluations in addition to qualitative ones.

The remainder of this paper is organized as follows. The next section provides different works made on semantic segmentation of omnidirectional images.

Section 3 introduces the experimental approach followed to achieve this work. Section 4 presents the results obtained and discusses them. Finally, Section 5 concludes the paper.

2 Related Work

Distinct works were carried out on semantic segmentation of omnidirectional images to compensate for the lack of algorithms dedicated to this type of data. In this section, we succinctly present the work done on fisheye images and on spherical images with their different representations.

2.1 Fisheye Images

Fisheye cameras have a field of view that can reach 180 degrees. Since CNNs for semantic segmentation are not designed for these images, and due to the unavailability of fisheye datasets with ground truth researchers worked on deformation of conventional images from Cityscapes or SYNTHIA [26], by applying a distortion to simulate the fisheye effect [27, 28, 11, 10]. The method used is described by $r_p = f\tan(r_f/f)$, which represents the mapping from the fisheye image point $P_f = (x_f, y_f)$ to the perspective image point $P_p = (x_p, y_p)$, where $r_p^2 = (x_p - u_{px})^2 + (y_p - u_{py})^2$ is the square distance between the image point P_p and the principal point $U_p = (u_{px}, u_{py})$ in the perspective image, and $r_f^2 = (x_f - u_{fx})^2 + (y_f - u_{fy})^2$ denotes the square distance between the image point P_f and the principal point $U_f = (u_{fx}, u_{fy})$ in the fisheye image. This only depends on a focal length; thus, several focal lengths were set to simulate different fisheye images with their corresponding annotations. Using the images resulting from this transformation, Deng et al. [11] proposed OPP-net based on an Overlapping Pyramid Pooling module, Saez et al. [27] proposed an adaptation of Efficient Residual Factorized Network (ERFNet) [25] to fisheye road images in order to achieve real-time semantic segmentation and tested it on real fisheye images but only qualitative results were exposed. Deng et al. [10] used the same method to achieve road scene semantic segmentation of fisheye surround-view cameras using restricted deformable convolution. The networks were trained on data from Cityscapes and SYNTHIA datasets and tested in real fisheye images.

2.2 Panoramic Images

Xu et al. [34] used synthetic images captured from SYNTHIA to create a dataset of panoramic images by stitching images taken from different directions. Using these images, the authors show that panoramic images improve segmentation results. Yang et al. [35] propose a panoramic annular semantic segmentation framework (PASS), such as the cited works on fisheye images authors made a data augmentation method by adding distortion to perspective images for the training set. And used normal CNNs after unfolding and partitioning the panoramic images.

2.3 Equirectangular Images

Equirectangular representation is the most popular projection for 360° images thanks to the simple transformation from spherical coordinates into planar coordinates. Classical CNNs designed for perspective images can be used for data under the equirectangular form. But spherical input suffers from distortion in polar regions. Different approaches were proposed to handle this issue. Monroy et al. [22] proposed SalNet360 where omnidirectional images were mapped to cubemap 6 faces projection and trained using normal CNNs to predict visual attention. However, artefacts are created when recombining the cubemap faces to omnidirectional image. Lai et al. [18] used semantic segmentation of equirectangular images to convert panoramic videos to normal perspective images. However for this task, highly accurate semantic segmentation was not required, in this work frame-based fully convolutional network FCN [21] was used. Su et al. [30] translated a planar CNN to process 360° images directly in the equirectangular projection for object detection. And in this publication [31] the same author proposed the kernel transformer network (KTN) to transfer convolution kernels from perspective images to equirectangular projection of 360° images efficiently for the same task. Tateno et al. [32] proposed a learning approach for equirectangular images using a distortion-aware deformable convolution filter for depth estimation from a single image, this approach was also demonstrated on 360° semantic segmentation.

2.4 Spherical representations

Because of distortions resulting from the equirectangular representation most recent work on this topic choose to work on the spherical presentation. Cohen et al. [7] developed spherical convolutions by replacing the translations in the plane by rotations of the sphere. Other work took advantage of the most accurate discretization of the sphere; the icosahedral spherical approximation. The discretization of the sphere presented by a spherical mesh generated by subdividing each face of a regular icosahedron into four equal triangles. Lee et al. [19] proposed an orientation-dependent kernel method regarding triangle faces, this method was demonstrated through classification, detection, and semantic segmentation. Zhang et al. [39] also addressed semantic segmentation on omnidirectional images using icosahedron spheres by proposing an orientation aware CNN framework. Jiang et al. [17] proposed UGSCNN to train spherical data mapped to an icosahedron mesh, by replacing conventional convolution kernels with linear combinations of learnable weighted operators. The last two contributions are the state-of-the-art of omnidirectional images semantic segmentation using spherical convolutions.

3 The Experimental Approach

To answer the questions addressed in the introduction, we carried out different experiments. We choose to use four CNNs developed for perspective images as

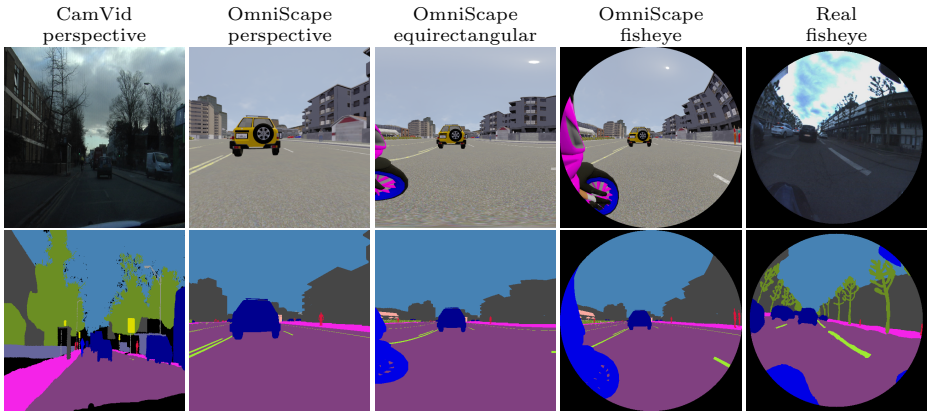


Fig. 1: Modalities used and corresponding semantic segmentation ground truth

well as UGSCNN which uses spherical convolutions. In the first experiment, we did a selection to choose the networks we will use in this study. The second experiment consists of a cross-modality experiment by training the four selected networks on real CamVid perspective images and fisheye, equirectangular and perspective OmniScape synthetic images. We tested the trained networks on all these modalities and also on our test set of fisheye images. In the last experiment, we trained UGSCNN on the same OmniScape equirectangular images used in the second experiment and tested it on the same modality with different resolutions. In all the experiments we use RGB images with 14 classes.

Table 1: Results of the networks selection (%)

	Mean global Acc.	mean IoU	Mean accuracies per-class														
			Void	Sky	Building	Fence	Other	Person	Pole	Road line	Road	Sidewalk	Vegetation	Two wheeled	Four wheeled	Wall	Traffic sign
FC-DenseNet56	91.8	60.3	46.4	96.7	90.5	75.8	63.3	58.5	41.3	97.0	97.9	**90.5**	88.3	73.6	89.1	76.4	55.1
FC-DenseNet67	**92.3**	54.4	47.7	**96.8**	92.2	**78.9**	**67.5**	62.7	**54.4**	96.9	**98.3**	88.6	87.7	73.9	**89.9**	77.1	**60.8**
FC-DenseNet103	**92.2**	**62.0**	49.4	96.7	91.7	78.5	65.4	57.2	46.3	97.4	98.2	90.2	**88.4**	72.7	89.7	77.3	55.0
MobileUNet [13]	87.6	48.9	37.0	93.6	87.1	73.4	53.2	33.6	15.0	96.5	96.8	83.2	83.0	62.6	80.1	66.4	34.6
PSPNet [40]	89.0	54.6	38.9	95.7	89.8	74.6	60.6	55.9	34.5	95.5	97.6	84.5	83.5	67.2	86.5	71.9	50.9
GCN [23]	90.7	56.2	42.1	96.3	90.5	71.5	52.2	53.6	40.5	96.0	97.9	89.7	86.0	66.0	83.6	74.1	49.4
FRRN	**91.9**	**61.8**	46.4	96.6	92.2	78.0	66.3	**64.9**	49.4	97.5	**98.3**	89.9	86.7	72.7	89.4	**77.6**	57.9
DeepLabV3 [4]	86.8	47.1	33.3	94.1	89.9	70.9	51.7	32.6	17.0	94.0	96.9	80.8	80.8	62.1	76.2	62.4	33.9
DeepLabV3+ [5]	89.3	53.2	39.7	95.1	89.5	72.6	53.8	45.4	33.0	94.4	97.8	86.6	87.1	64.2	84.0	68.5	45.5
RefineNet	91.2	**59.3**	42.9	96.0	**92.5**	75.5	60.6	57.0	39.8	**97.7**	98.1	89.1	87.4	71.0	86.3	74.5	51.9
AdapNet [33]	87.3	47.9	38.6	96.7	89.2	71.9	52.8	26.5	18.3	96.3	96.2	78.3	80.1	61.0	76.8	65.6	34.5
DenseASPP [36]	87.9	50.6	39.5	91.4	90.5	71.4	54.9	41.3	23.9	94.8	97.6	83.1	82.2	65.2	78.1	67.7	37.4
BiSeNet [38]	90.3	55.1	40.2	95.9	90.6	74.6	53.7	47.0	24.9	96.9	97.9	88.2	87.6	65.8	85.3	70.6	50.6
SegNet	**92.0**	**61.8**	**50.1**	96.2	92.1	78.5	66.5	59.3	46.3	97.5	98.0	89.5	88.0	**74.3**	89.0	76.6	57.0

3.1 Networks selection

The goal of this experiment is to choose four networks we will use in the cross-modality. To choose these networks we made a selection using CamVid Dataset among 11 networks representing the state of the art on semantic segmentation of perspective images. We trained and tested all the networks on same sets of 512x512 CamVid images. We used 700 images, 420 in the training set, 112 in the validation set and the remaining 168 images in the test set. CamVid dataset offers perspective images with per-pixel semantic segmentation of over 700 images. The images are segmented into 32 object classes. We mapped similar classes into 14 to have the same classes present in OmniScape. Fig. 1 shows a CamVid image with ground truth. The results of this first selection are presented in Table 1. We can notice that all networks are quite similar in general. However, the four networks which give the best Intersection over Union (IoU) score with good average accuracy are Fully Convolutional DenseNet, Full-Resolution Residual Network, SegNet, and RefineNet. For Fully Convolutional DenseNet network, we chose to use just the architecture built from 103 convolutional layers for the next experiment. In the following we present a brief overview on each of the four chosen networks.

Fully Convolutional DenseNet [15] This network is an adaptation of DenseNets for semantic segmentation. It is a U-Net architecture where the convolutional layers are replaced with dense blocks. Each convolution layer is then directly connected to every other layer. **Full-Resolution Residual Network** [24] This network combines two distinct processing streams. One stream undergoes a sequence of pooling operations and is responsible for understanding large-scale relationships of the elements in the image. The second stream carries feature maps at the full image resolution, giving a precise adherence to boundaries. The pooling operations in the first stream act like residual units for the second, and carry high level information over the network. **SegNet** [1] This network consists of an encoder-decoder layer followed by a pixel-wise classification layer. The architecture of the encoder layer is identical to the VGG16 network. Each encoder is one or more convolutional layers. This layer contains batch normalization, a ReLU non-linearity, a non-overlapping maxpooling, and sub-sampling. **RefineNet** [20] This network is considered as a generic multi-path refinement network which uses long range residual connections to enable high resolution prediction by exploiting all the information available in the down-sampling process. Like this by using fine-grained features from earlier convolution, the deeper layers that capture high level semantic features can be directly refined.

3.2 Cross-modality experiment

In this experiment we used 700 captures from OmniScape, 700 images from CamVid and 15 images of our real fisheye images test set. **The OmniScape Dataset** provides synthetic omnidirectional images namely, 360° equirectangular, fisheye and catadioptric stereo RGB images from the two front sides of a motorcycle with semantic segmentation and depth map ground truth. It provides also the tools to generate omnidirectional images using intrinsic parameters of

Table 2: Image sets and networks used in the cross-modality

Training Sets	Testing sets	Networks
- CamVid Perspective images - OmniScape Perspective images - OmniScape Fisheye images - OmniScape Equirectangular images	- CamVid Perspective images - OmniScape Perspective images - OmniScape Fisheye images - OmniScape Equirectangular images - Real Fisheye images	- FC-DenseNet103 - SegNet - FRRN - RefineNet

an omnidirectional camera and a 360º cubemap image. We exploit this tool to generate omnidirectional images using the same intrinsic parameters of our calibrated fisheye camera. The images in OmniScape are annotated into 14 classes automatically since it is synthetic data. For Equirectangular representation, we crop the images to keep just 180 degrees which represents the front side, so all modalities can be compared to each other. Fig. 1 shows OmniScape different modalities used with semantic segmentation ground truth. **Our test set** contains real fisheye images captured using the same disposition used in the OmniScape dataset; Stereo fisheye cameras placed in the two front sides of a motorcycle. We annotated 15 different images into 14 classes like the OmniScape dataset, using the open source tool for annotation PixelAnnotationTool [2]. Fig. 1 shows an example of an image from this set with ground truth. We split the 700 images of each modality like a standard cross validation problem into 3 sets: a training set of 420 images, a validation set of 112 images, and a test set of 168 images. We use the four chosen networks to train on OmniScape images using fisheye, perspective and 180º equirectangular images and also CamVid images. Then we test all the trained networks on the three modalities of OmniScape images, CamVid images and our test set of fisheye real images annotated manually. This leads to 16 training processes and 20 test processes for each of the four training sets. The class Void in CamVid represents far objects that are undefined, and in OmniScape dataset, it represents the dark space surrounding the fisheye image. In this experiment we drop this class and we do not take it into account in the calculation of the scores because it does not represent an information. In Table 2 are listed the training and testing sets along with the networks used in the cross-modality.

3.3 Comparison with spherical CNNs

In this experiment, we trained UGSCNN on the same OmniScape equirectangular images used in the second experiment. This time instead of cropping the images we replaced the pixels representing the 180 degrees of the back side by zero and treated it like a new class Void. This class is not evaluated for performances. We performed this experiment in three resolutions 5, 7, and 8 as shown in Fig. 2. Since the resolution of the images used is 512x1024 we settle for these resolutions. We used in this experiment just RGB, without depth map since the depth map was not used by the other networks. The network is trained with a batch size of 16 for resolutions 5 and 7 and a batch size of 8 for resolution 8. We

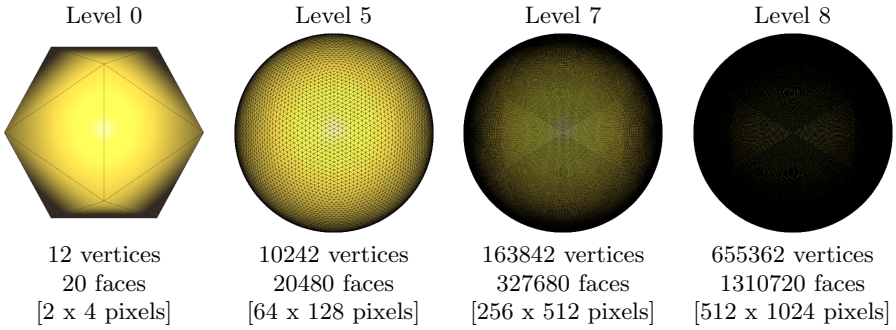


Fig. 2: Icosahedral subdivision levels, the corresponding equirectangular pixel resolutions and number of elements.

use like in [17] the weighted cross-entropy loss for training, and zero weight for the dropped class Void. To display qualitative results, we unwrap the sphere using the UV mapping process. The equirectangular images are regenerated using the following for any point P on the sphere:

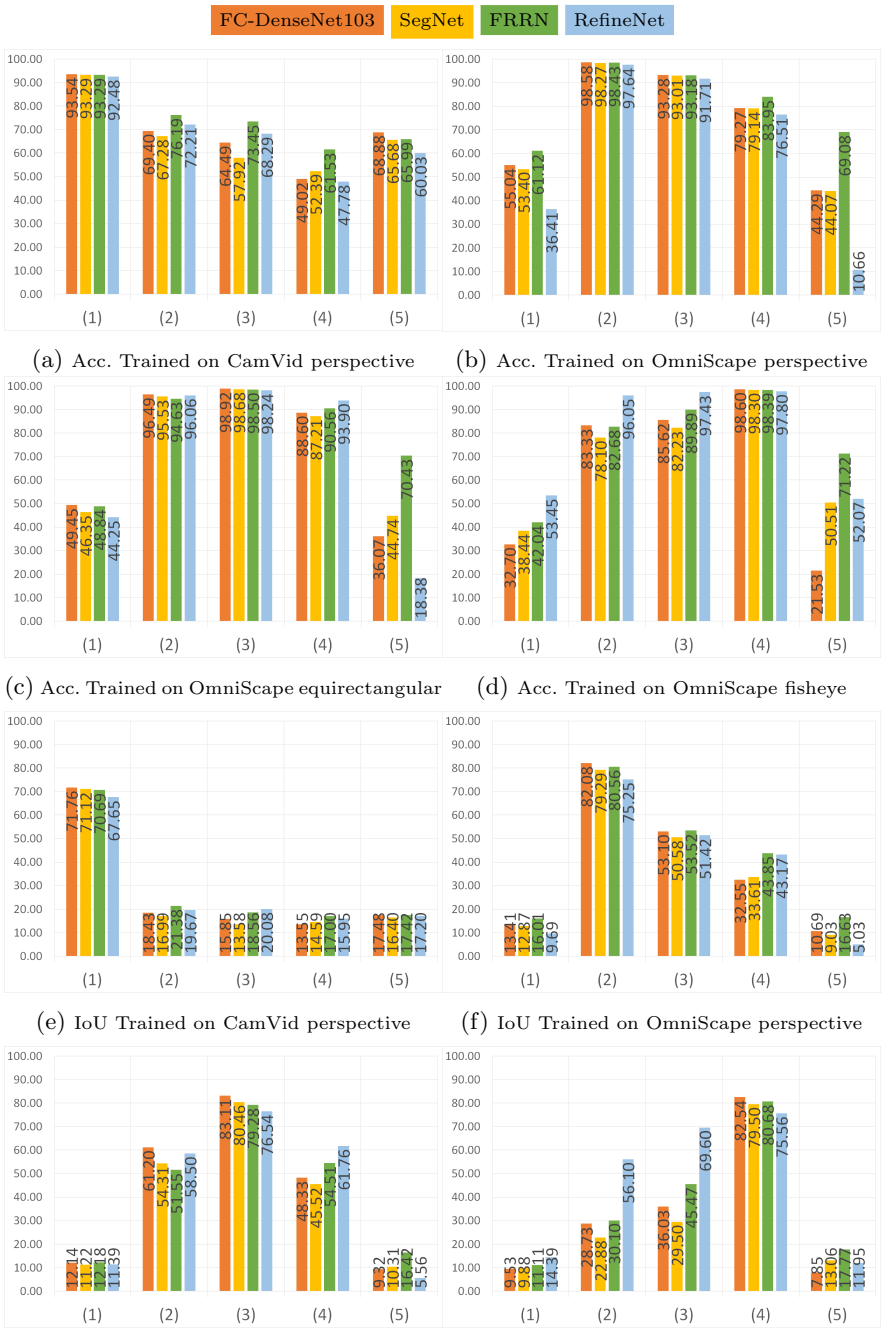
$$u = 0.5 + \frac{\arctan2\left(d_z, d_x\right)}{2\pi} \qquad v = 0.5 - \frac{\arcsin\left(d_y\right)}{\pi} \tag{1}$$

where $(\hat{u, v})$ are the coordinates in the equirectangular image in the range [0, 1] and $\hat{d}$ the unit vector from P to the sphere's origin. Fig. 6 shows examples of unwrapped equirectangular images from a sphere with different resolutions.

UGSCNN is an orientation-aware method. In this network, the convolution kernel is replaced by linear combinations of differential operators that are weighted by learnable parameters using standard back-propagation. The operators are estimated on unstructured grids.

4 Results and Discussions

In this section, we present the results of the cross-modality experiment and the comparison with spherical CNNs. We discuss and give quantitative results as well as qualitative ones. We answer the questions addressed in the introduction by analyzing the results given by networks trained on omnidirectional images and those trained on perspective images. And finally, we make a comparison between the combination network/training-set which gives the best results on equirectangular images in the first experiment and UGSCC trained and tested on the same data with different resolutions. Fig. 3 represents an overview of results obtained in the cross-modality experiment using clustered column. It resumes all the results obtained by 80 testing processes. As a first remark, we can see that always best results are obtained when the modality does not change between training and testing processes. And the four networks are very sensitive to texture changes. We see that when the environment changes the performance deteriorates drastically.



(a) Acc. Trained on CamVid perspective (b) Acc. Trained on OmniScape perspective

(c) Acc. Trained on OmniScape equirectangular (d) Acc. Trained on OmniScape fisheye

(e) IoU Trained on CamVid perspective (f) IoU Trained on OmniScape perspective

(g) IoU Trained on OmniScape equirectangular (h) IoU Trained on OmniScape fisheye

(1) Camvid perspective, (2) OmniScape perspective, (3) OmniScape Equirectangular, (4) OmniScape fisheye, (5) Real fisheye

Fig. 3: Per test set mean accuracies and IoU (%)

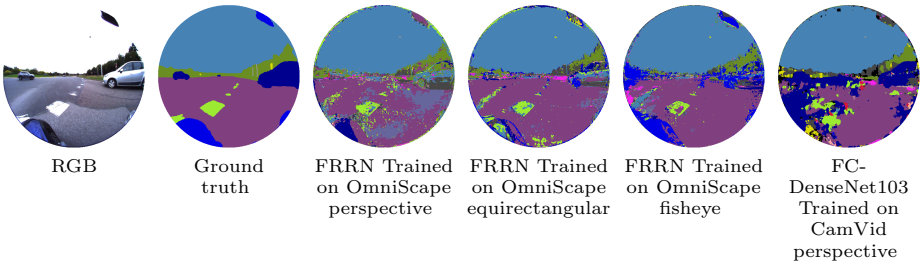


Fig. 4: Qualitative results on a real fisheye image using networks given best IoU for each modality

4.1 Omnidirectional images

The four networks when trained on fisheye images or equirectangular images and tested on the same modalities give a mean accuracy not less than 97.80% and a mean IoU higher than 75.56% without exception. It shows clearly that networks designed for perspective images when trained on omnidirectional images give good results. This answers the first question in the introduction, we don't need to adapt networks for omnidirectional images. Just training them on omnidirectional images is good enough to achieve results similar to those obtained by training and testing on perspective images.

4.2 Real fisheye images

Results obtained for real fisheye images are unexpectedly very poor, the highest obtained IoU being 17.77%. However, best results are reached when OmniScape fisheye images were used in training. This can be explained by the fact that in CamVid, textures match with the real fisheye images but the geometry does not. And in OmniScape fisheye images it is the opposite, the geometry is similar, since we use the same intrinsic parameters of our real fisheye camera, but textures are very different between OmniScape and real images. On the other hand, results obtained when training and testing on OmniScape images all modalities combined are better. We can conclude that both texture and geometry are important. But the geometry slightly outweighs the texture in this case. If real fisheye images or more realistic synthetic fisheye images were used in training, results could be much better. Fig. 4 shows qualitative results obtained by each time the best network in each modality. It is worth noting that Accuracy and IoU are computed without taking in account the surrounding black area in fisheye images. We consider just the part which contains the information. FRRN with OmniScape fisheye images gives the best results when testing on real fisheye images. However, it is not the fastest in terms of computation time as shown in Table 3. FC-DenseNet103 and FRRN when using CamVid perspective images are fairly close to each other but FRRN is faster. In a real-time application, FRRN is to be preferred in this case.

Table 3: Average runtime of the selected networks using NVIDIA Quadro P3200

Network	Runtime (ms)
SegNet	263.4
RefineNet	271.6
FRRN	349.6
FC-DenseNet103	795.2

Table 4: Average runtime of UGSCNN and FC-DenseNet103 using NVIDIA Tesla V100 SXM2

Network	Runtime (ms)
UGSCNN level 5	58.1
FC-DenseNet103	155.1
UGSCNN level 7	879.3
UGSCNN level 9	3565.1

Table 5: Networks with best mean IoU (%) in the cross-modality experiment

		Testing					
		Perspective		Equirectangular		Fisheye	
Training	Perspective	FC-DenseNet103	82.08	FRRN	53.52	FRRN	43.85
	Equirectangular	FC-DenseNet103	61.20	FC-DenseNet103	83.11	RefineNet	61.76
	Fisheye	RefineNet	56.10	RefineNet	69.60	FC-DenseNet103	82.54

4.3 OmniScape images

In these images the only difference is the camera itself, the same scene is captured by three cameras respectively perspective, fisheye, and 360° equirectangular. This configuration allows us to make a fair comparison between all these modalities. When trained on perspective images all the four networks achieve the best scores for perspective images, but when tested on omnidirectional images we lose 29.59% for equirectangular images and 38.69% for fisheye images in mean IoU when we consider the best network each time. On the other hand, we notice that networks trained on equirectangular images are quite robust for both fisheye and perspective because we lose just 20.78% for fisheye and 20.88% for perspective. Otherwise, networks trained on fisheye loose just 13.51% when tested on equirectangular. This shows that networks, when trained on omnidirectional images, are more robust and can learn a universal representation more than when trained on perspective images. We can notice that when the testing and training modalities are similar FC-DenseNet103 is slightly better than the others. When the modalities change FRRN is better when training on perspective images and RefineNet is better when training on fisheye images. Due to space limitation, we present just one case of qualitative results comparison, in Fig. 5 predicted equirectangular and perspective images when using RefineNet trained on fisheye images and FC-DenseNet103 trained on perspective and equirectangular are shown. Table 5 shows the best results for all the nine combinations. We can see that RefineNet trained on fisheye images is quite robust when tested on other modalities. There are small differences almost invisible especially for classes with a low-class weight across the sets.

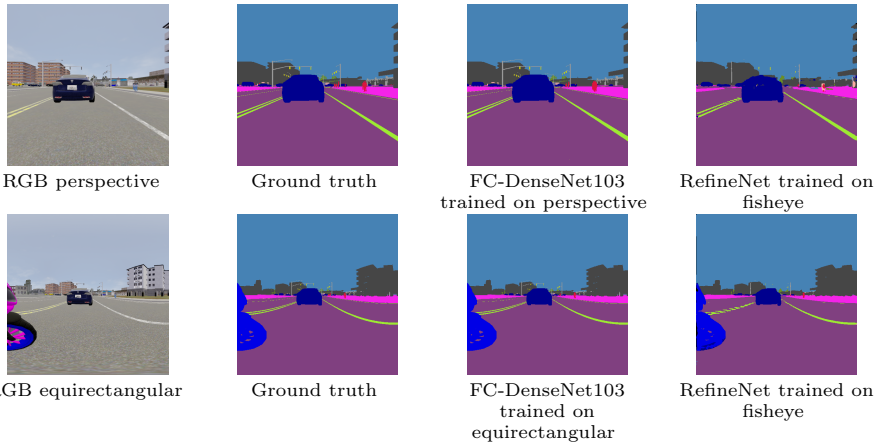


Fig. 5: Qualitative results of RefineNet trained on fisheye and tested on perspective and equirectangular

4.4 Spherical images

The motivation behind this experiment is to know if spherical convolution gives better results than networks used in the second experiment when tested on equirectangular images. We saw in the previous experiment that FC-DenseNet103 gives the best results when trained and tested on equirectangular. This is our baseline. As explained before we used three resolutions. But the resolution which is comparable to images we used in the previous experiment is level 8 because we use 512x512 images, as shown in Fig. 2. We will focus especially on level 7 and level 8 since results obtained using level 5 are not usable, this can be verified by visualizing the images in Fig. 6. When mapping the equirectangular 512x1024 images in a sphere level 5, we lose a lot of data and we can see that the structure of objects in the images is not conserved even if the scene looks similar. In Table 6 mean accuracies are listed and in Table 7 mean IoU. FC-denseNet103 gives better results in terms of mean accuracy and IoU, a big gap can be observed between the two networks. This is due to the fact that UGSCNN predicts a lot of false negatives. For example, the class Four-wheeled has the highest accuracy but a very low IoU. This is because UGSCNN predicts four wheeled pixels quite good in the images but it also predicts a lot of other classes like four wheeled. In Fig. 6 qualitative results are presented, we can notice the presence of blue stains (the blue color which represents four wheeled in the semantic segmentation) in the UGSCNN level 8 predicted image which explains this difference between accuracy and IoU, the same problem is present also for level 5. UGSCNN is considered as one of state-of-the-art network for spherical CNNs but it is still far behind networks that use normal convolution and is extremely slow in terms of runtime. Table 4 shows the average runtime comparison. FC-DenseNet103 is 23 times faster than UGSCNN level 8.

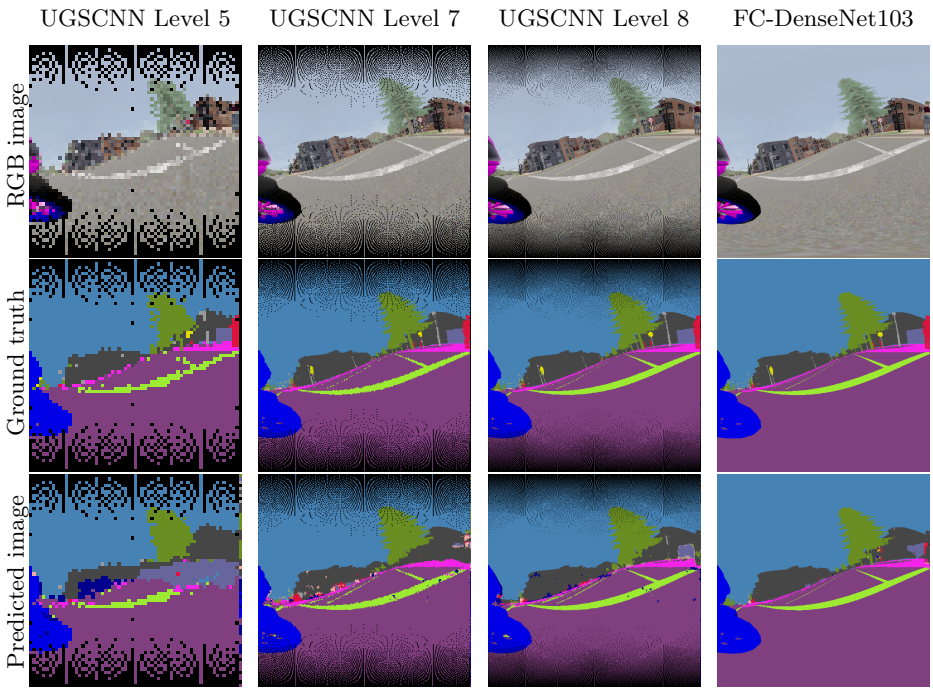


Fig. 6: Qualitative results for UGSCNN and FC-DenseNet103

4.5 Summary

To summarize all the experiments conducted in this work, we can say that semantic segmentation networks made for perspective images give good results and are more robust when trained on omnidirectional images. They are able to learn a universal representation and achieve better results on all modalities than if trained on perspective images. Finally, we made a comparison between a network that uses spherical CNNs and a network with normal convolutions using equirectangular images. Working on the sphere is very greedy in terms of computation time and memory but does not necessarily give better results. Even if we can consider that it is two different modalities but the input remains the same; an equirectangular image.

5 Conclusion and Future Work

This paper takes stock of progress made on semantic segmentation of omnidirectional images. We presented a comparative study of semantic segmentation using equirectangular, fisheye, and perspective images, from real and synthetic datasets. By comparing different networks of semantic segmentation, we proved that networks developed for perspective images when trained on omnidirectional images give good results and they become more robust against modality changes.

Table 6: Mean accuracies of UGSCNN and FC-DenseNet103 (%)

	Mean global Acc.	Mean accuracies per-class													
		Sky	Building	Fence	Other	Person	Pole	Road line	Road	Sidewalk	Vegetation	Two wheeled	Four wheeled	Wall	Traffic sign
FC-DenseNet103	**98.9**	**99.4**	**96.2**	39.1	50.4	**31.7**	42.1	90.8	**99.8**	**95.7**	86.7	**99.6**	74.4	81.4	29.5
UGSCNN level 5	72.0	96.9	87.4	49.7	40.1	11.7	38.3	90.6	97.0	90.9	88.7	98.5	87.7	84.1	46.9
UGSCNN level 7	78.4	96.8	92.7	51.9	54.1	22.6	52.7	95.2	98.0	92.5	91.7	98.4	91.6	**91.5**	**67.3**
UGSCNN level 8	79.5	97.1	94.7	**57.0**	**50.7**	26.2	**56.4**	**96.2**	98.0	**95.7**	**92.2**	98.5	**93.0**	90.7	67.1

Table 7: Mean IoU of UGSCNN and FC-DenseNet103 (%)

	Mean IoU	Mean IoU per-class													
		Sky	Building	Fence	Other	Person	Pole	Road line	Road	Sidewalk	Vegetation	Two wheeled	Four wheeled	Wall	Traffic sign
FC-DenseNet103	**72.3**	**98.9**	**91.6**	32.6	**44.1**	**27.0**	**38.4**	**87.9**	**99.5**	**90.9**	**76.6**	**99.3**	**67.5**	72.8	**27.4**
UGSCNN level 5	41.3	96.4	20.1	31.4	1.6	7.6	22.8	73.6	95.3	67.1	15.1	61.0	2.7	63.0	20.5
UGSCNN level 7	43.7	94.5	16.6	**37.6**	0.6	0.3	7.3	77.9	97.5	79.0	36.6	15.0	57.7	82.3	9.3
UGSCNN level 8	42.9	96.6	45.7	35.3	0.7	7.1	15.2	78.6	97.7	80.0	10.6	17.5	6.1	**86.0**	23.4

We also made a comparison using equirectangular images with both normal convolution and spherical. The experiment shows that normal convolution is better. As we noticed that networks used are sensitive to textures and environment changes, one solution can be to use networks performing image to image translation like pix2pix [14] to generate more realistic images using OmniScape dataset since we lack datasets of real omnidirectional images with ground truth. Recently, WoodScape dataset [37] of real fisheye images with pixel-wise ground truth was published but not yet available. We can also explore data fusion methods in order to learn a better universal representation that can be robust against modality and environment changes. Ideally, a network able to learn shapes and geometry of objects regardless of texture and position on the omnidirectional image would be more adequate for omnidirectional images.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (Dec 2017)
2. Bréhéret, A.: Pixel Annotation Tool. github.com/abreheret/PixelAnnotationTool (2017)
3. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters **30**, 88–97 (2009)
4. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017)
5. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018)
6. Cohen, T., Welling, M.: Group equivariant convolutional networks. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 2990–2999. PMLR, New York, New York, USA (20–22 Jun 2016)
7. Cohen, T.S., Geiger, M., Köhler, J., Welling, M.: Spherical cnns. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
10. Deng, L., Yang, M., Li, H., Li, T., Hu, B., Wang, C.: Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. IEEE Transactions on Intelligent Transportation Systems pp. 1–13 (2019)
11. Deng, L., Yang, M., Qian, Y., Wang, C., Wang, B.: Cnn based semantic segmentation for urban traffic scenes using fisheye camera. In: 2017 IEEE Intelligent Vehicles Symposium (IV). pp. 231–236 (June 2017)
12. Eder, M., Shvets, M., Lim, J., Frahm, J.M.: Tangent images for mitigating spherical distortion (2019)
13. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR **abs/1704.04861** (2017)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. arxiv (2016)
15. Jegou, S., Drozdzal, M., Vazquez, D., Romero, A., Bengio, Y.: The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (July 2017)
16. Jeon, Y., Kim, J.: Active convolution: Learning the shape of convolution for image classification. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
17. Jiang, C.M., Huang, J., Kashinath, K., Prabhat, Marcus, P., Niessner, M.: Spherical CNNs on unstructured grids. In: International Conference on Learning Representations (2019)

18. Lai, W., Huang, Y., Joshi, N., Buehler, C., Yang, M., Kang, S.B.: Semantic-driven generation of hyperlapse from 360 degree video. IEEE Transactions on Visualization and Computer Graphics **24**(9), 2610–2621 (Sep 2018)
19. Lee, Y., Jeong, J., Yun, J., Cho, W., Yoon, K.J.: Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
20. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
21. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3431–3440 (June 2015)
22. Monroy, R., Lutz, S., Chalasani, T., Smolic, A.: Salnet360: Saliency maps for omnidirectional images with cnn. Signal Processing: Image Communication **69**, 26 – 34 (2018), salient360: Visual attention modeling for 360° Images
23. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters — improve semantic segmentation by global convolutional network. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1743–1751 (July 2017)
24. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
25. Romera, E., Álvarez, J.M., Bergasa, L.M., Arroyo, R.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Transactions on Intelligent Transportation Systems **19**(1), 263–272 (Jan 2018)
26. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3234–3243 (June 2016)
27. Saez, A., Bergasa, L., López-Guillén, E., Romera, E., Tradacete, M., Gómez-Huélamo, C., del Egido, J.: Real-time semantic segmentation for fisheye urban driving images based on erfnet. Sensors **19**(3), 503 (Jan 2019)
28. Sáez, Á., Bergasa, L.M., Romera, E., Guillén, M.E.L., Barea, R., Sanz, R.: Cnn-based fisheye image real-time semantic segmentation. 2018 IEEE Intelligent Vehicles Symposium (IV) pp. 1039–1044 (2018)
29. Sekkat, A.R., Dupuis, Y., Vasseur, P., Honeine, P.: The omniscape dataset. In: 2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - June 4, 2020. IEEE (2020)
30. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360°imagery. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 529–539. Curran Associates, Inc. (2017)
31. Su, Y.C., Grauman, K.: Kernel transformer networks for compact spherical convolution. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
32. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: The European Conference on Computer Vision (ECCV) (September 2018)
33. Valada, A., Vertens, J., Dhall, A., Burgard, W.: Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In: 2017 IEEE International Conference on Robotics and Automation (ICRA). pp. 4644–4651 (May 2017)

34. Xu, Y., Wang, K., Yang, K., Sun, D., Fu, J.: Semantic segmentation of panoramic images using a synthetic dataset. In: Dijk, J. (ed.) Artificial Intelligence and Machine Learning in Defense Applications. vol. 11169, pp. 90 – 104. International Society for Optics and Photonics, SPIE (2019)
35. Yang, K., Hu, X., Bergasa, L.M., Romera, E., Huang, X., Sun, D., Wang, K.: Can we pass beyond the field of view? panoramic annular semantic segmentation for real-world surrounding perception. In: 2019 IEEE Intelligent Vehicles Symposium (IV). pp. 446–453 (June 2019)
36. Yang, M., Yu, K., Zhang, C., Li, Z., Yang, K.: Denseaspp for semantic segmentation in street scenes. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
37. Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricar, M., Milz, S., Simon, M., Amende, K., Witt, C., Rashed, H., Chennupati, S., Nayak, S., Mansoor, S., Perrotton, X., Perez, P.: Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
38. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: The European Conference on Computer Vision (ECCV) (September 2018)
39. Zhang, C., Liwicki, S., Smith, W., Cipolla, R.: Orientation-aware semantic segmentation on icosahedron spheres. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
40. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)