Panoramic Depth and Semantic Estimation with Frequency and Distortion aware Convolutions

Bruno Berenguel-Baeta* and Jesus Bermudez-Cameo and Jose J. Guerrero

***Abstract*—Omnidirectional images reveal advantages when addressing the understanding of the environment due to the 360-degree contextual information. However, the inherent characteristics of the omnidirectional images add additional problems to obtain an accurate detection and segmentation of objects or a good depth estimation. To overcome these problems, we exploit convolutions in the frequency domain, obtaining a wider receptive field in each convolutional layer, and convolutions in the equirectangular projection, to cope with the image distortion. Both convolutions allow to leverage the whole context information from omnidirectional images. Our experiments show that our proposal has better performance on non-gravity-oriented panoramas than state of the art methods and similar performance on oriented panoramas as specific state of the art methods for semantic segmentation and for monocular depth estimation, outperforming the sole other method which provides both tasks. Code is publicly available in https://github.com/Sbrunoberenguel/FreDSNet**

I. Introduction

In recent years, the research community has put a great effort into the 3D scene understanding problem from images [1]. Among the different branches in the scene understanding field, object detection and segmentation [2] has great interest for interaction while depth estimation [3] is more interesting for 3D mapping and navigation.

Previous approaches to indoor scene understanding rely on using conventional cameras with classic algorithms (i.e. ORB-SLAM [4] for 3D mapping and localization). With the rise of deep learning approaches [5], several solutions for depth [6], [7] and semantic segmentation [8] from perspective images appeared. Even though these solutions provide good results, conventional cameras still lack one key feature in navigation: awareness of the surroundings. Conventional cameras provide a limited field of view which makes it difficult to obtain a complete knowledge of the environment.

Omnidirectional cameras, such as fish-eye or panoramic cameras, provide a better understanding of the environment in a single image. With a field of view up to 360 degrees, these cameras encode the whole environment's information into a single image. However, we find some challenges that we do not find on conventional cameras. The first challenge is the image distortion of these omnidirectional cameras. Due to the wide field of view, the projection model of these cameras carry heavy distortion in some areas of the image (i.e. in equirectangular panoramas, the top and bottom part of the images are heavily distorted since, in the extreme, one pixel is stretched in the whole image width). The second challenge is the few number of labelled datasets for general or specific purposes. Since these images have only being used in the last decade and it is difficult to manually annotate the images due to the distortion, there are few datasets and the ones that can be found are not very large.

In this work, we aim to cope with these scene understanding challenges. We present a neural network that jointly provides semantic segmentation and monocular depth estimation of indoor environments from a single equirectangular panorama (see Fig. 1). We propose the use of the Fast Fourier convolution (FFC) [9] to leverage the wider receptive field of these convolutions. We also use convolutions adapted to the equirectangular distortion [10] to take advantage of the wide field of view of 360 panoramas. Besides, we propose a joint training of semantic segmentation and depth, where each task can benefit from the other. We further explore the use of different loss functions for semantic segmentation and optimization processes to compute the hyper-parameters for the global loss function. From the scarcity of labelled data, we present a new tool for a better panoramic image stitching and the possibility of obtaining semantic segmentation maps in equirectangular projection from the Matterport3D dataset [11].

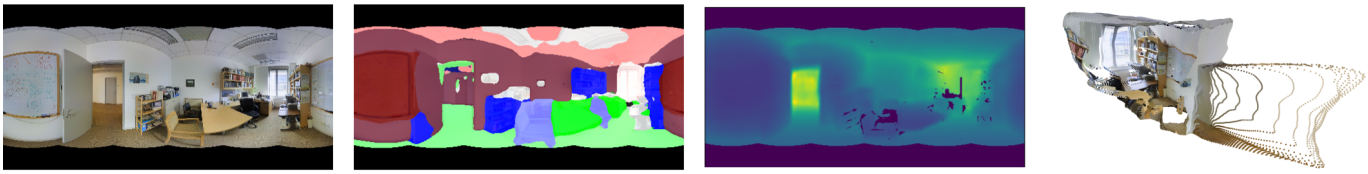
The main contributions of this work can be summarized as:

1) We evaluate the robustness and adaptability of several learning-based solutions for monocular depth estimation and semantic segmentation on non-gravity-oriented panoramas.
2) We jointly exploit convolutions adapted to the equirectangular projection, obtaining a better distortion management in the equirectangular panorama, and the fast Fourier convolution, obtaining a wider receptive field in early layers of the network increasing the context information for indoor scene understanding.
3) We propose a joint training of monocular depth and semantic segmentation, leveraging the similarities between both tasks to obtain better predictions.

II. Related Work

Among the 3D scene understanding topics, the semantic segmentation and monocular depth estimation problems have attracted the attention of researchers in the last years. Besides, with the use of omnidirectional images, we can find several approaches to cope with the distortion and leverage context information.

*Corresponding author: berenguel@unizar.es
Bruno Berenguel-Baeta ORCID: 0000-0003-2674-4844
All authors are with Instituto de Investigacion Ingenieria Aragon, Universidad Zaragoza, Zaragoza, Spain



(a) (b) (c) (d)

Fig. 1: Overview of our proposal. From a single RGB panorama (a), we make a semantic segmentation (b) and estimate a depth map (c) of an indoor environment. With this information we are able to reconstruct in 3D the whole environment (d).

Semantic segmentation The semantic segmentation on perspective images is a well-studied field. We can find many works on object detection [12], semantic segmentation [13], [14] or both tasks [15] from perspective cameras. These approaches provide good results but lack context information. On the other hand, the use of omnidirectional images provides a better understanding of the surroundings and new approaches are appearing. We can find works on object detection [16] or segmentation [17] or both at the same time [18] which use convolutional neural networks, attention modules [19], [20] or transformers [21], [22]. Since omnidirectional images present heavy distortions (e.g. in spherical projections, like equirectangular images, this distortion is more accentuated in the mapping of the poles) these kinds of images are difficult to manually annotate. Nevertheless, due to the wide field of view of these images (e.g. in the spherical projection, we can see all the surroundings in a single image), the use of omnidirectional images in semantic segmentation is an active field of study since we can obtain a complete semantic understanding of the environment from a single image.

Depth estimation Monocular depth estimation is a research topic that has been in the spotlight in recent years. Recent approaches rely on deep learning methods, such as [23], which proposes convolutions adapted to the camera calibration, [24] propose relative depth maps and [25] estimate the camera calibration and depth maps at the same time, all in perspective cameras. On omnidirectional cameras, first approaches used convolutional networks [26]. New ones combine convolutional networks with: recurrent networks [27], attention modules [19], [28] or transformers [29], [22]. The use of twin networks has also been studied in [30], [31] as well as different fusion methods [32]. Each work presents particular approaches for monocular depth estimation, being an open field of study with great interest and many applications.

Network architecture Previous works on semantic segmentation or depth estimation rely on convolutional encoder-decoder architectures with some recurrent [27] or attention mechanism [19] as hidden representation of the environment. These approaches of encoder-decoder architecture which rely on standard convolutions [26] suffer from slow growth of the receptive field of the convolutions, losing general context information. More recent approaches overcome this problem decomposing the panoramic image into several gnomic projections to take advantage of standard convolutions [17] or visual transformers [33], [34], using a spherical geometric positional encoding for feature fusion.

Our method is inspired by works that try to adapt convolutional neural networks [5] to the particular distortion of equirectangular panoramas [21]. In particular, we propose an encoder-decoder architecture with convolutions adapted to spherical distortion, EquiConvs [10], as well as convolutions in the frequency domain with a higher receptive field, Fast Fourier convolutions (FFC) [9]. With distortion-aware convolutions and a higher receptive field, we are able to obtain more context information of the environment directly from a single panorama. We believe that learning directly from the spherical model provides continuous context information, which is lost in the decomposition of the panorama in patches of transformed-based networks. Besides, adapting the network to the distortion allows a more robust behaviour against more general conditions (i.e. panoramas under different orientations).

III. Monocular Depth and Semantic Segmentation

In this article we build upon the architecture presented in [35]. We present an encoder-decoder architecture with Resnet-50 [36] as initial feature extractor and separate branches for depth estimation and semantic segmentation, as seen in Fig.2. We use multi-resolution encoding and decoding, in order to obtain a multi-scale features, and the use of weighted skip connections between encoder and decoder. Each branch takes intermediate feature maps from the decoder, obtaining multi-scale information. The main novelties of the architecture are the building blocks used, combining frequency and spatial convolutions, and the multi-branch decoder designed for the joint training of both tasks.

As main difference with the previous work, we change the standard convolutions by convolutions adapted to the equirectangular projection in parallel with the fast Fourier convolutions. Furthermore, we also make an extended study of the loss function for training and the adjustment of the hyper-parameters. In this section we present these novelties.

A. Equirectangular Convolutions

Kernels are the keystone of convolutional neural networks (CNN). Conventional kernels are square filters that go through images or feature maps obtaining new feature maps. Previous works addressed the problem: what happens if we change the shape of the kernel? The work [37] was the first to present a learned deformable kernel, improving the performance of CNN on several tasks. In this work, we also propose the use of deformable kernels, but in our case we are not learning them. We use the equirectangular projection to compute the

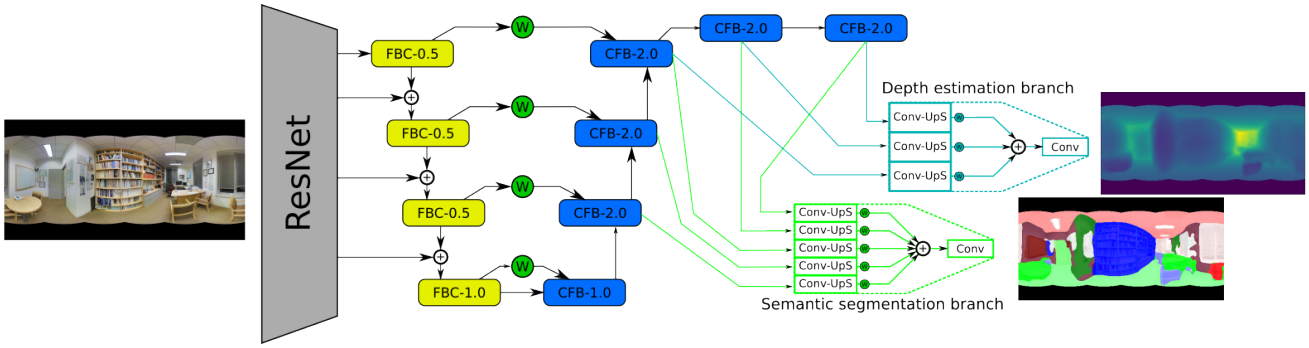


Fig. 2: Network architecture. The encoder part is formed by a feature extractor (ResNet50) and four encoder blocks. The decoder part is formed by six decoding blocks and two branches that predict depth and semantic segmentation.

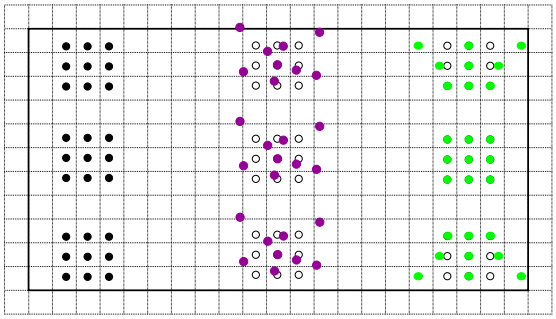


Fig. 3: Equirectangular panorama with padding and convolutions on different places. From left to right: Standard convolution, learned deformable convolution [37], Equirectangular convolution [10]. In empty black circles are where the standard convolutions would be located.

kernel deformation adapting our network to the distortion of equirectangular panoramas. Figure 3 shows the differences among standard, deformable and equirectangular kernels.

The convolutional kernels that we use in this work are called *EquiConvs* and were first presented by [10]. Taking advantage of the closed form of the equirectangular projection, we can deform the kernel to adapt its geometry to the equirectangular distortion. This deformation comes from computing a set of offsets for the kernel in each pixel of the image or feature map.

In a similar manner as [10], we define a kernel of resolution $r_w \times r_h$ as: $\hat{\mathbf{k_{ij}}} = [i, j, d]$, where i, j are in the range $\left[-\frac{r-1}{2}, \frac{r-1}{2}\right]$ and d is the distance from the center of the sphere to the kernel.

Defined the kernel as a plane tangent to the sphere, we rotate the kernel to match its center with the pixel coordinates we want to apply the convolution on as: $\mathbf{k_{ij}} = [x_{ij}, y_{ij}, z_{ij}] = \mathcal{R}_{ot}\ \hat{\mathbf{k_{ij}}}/|\hat{\mathbf{k_{ij}}}|$, where $\mathcal{R}_{ot}$ is the rotation matrix to the pixel where we apply the convolution. Once defined the kernel in the correct position, we use the back projection model to define the kernel back into the equirectangular image domain.

$$\phi_{ij} = \arctan\left(\frac{x_{ij}}{z_{ij}}\right);\ \ \theta_{ij} = \arcsin\left(y_{ij}\right) \tag{1}$$

First, we compute the spherical coordinates of the kernel's projecting rays (see eq. 1) and then we compute the pixel coordinates as:

$$u_{ij} = \left(\frac{\phi_{ij}}{2\pi} + 0.5\right) W;\ \ v_{ij} = \left(-\frac{\theta_{ij}}{\pi} + 0.5\right) H, \tag{2}$$

where (W, H) is the panorama width and height and (u_{ij}, v_{ij}) are the pixel coordinates of the kernel in the image.

B. Loss function

We introduce the loss function used for training, based in the prior work [35]. We define each individual loss and present our combined loss function. Semantic segmentation and depth estimation have common characteristics that can benefit from each other [38](i.e. object edges are clear on the semantic segmentation map and define a discontinuity in the depth map; room layouts define 3D planes in the depth map and can aid the detection of the structural elements in the segmentation map). We want to take advantage of these similarities by making a joint training where the semantic segmentation and the depth estimation can be jointly predicted.

For the semantic segmentation loss L_{Seg} , we use the standard *Cross Entropy Loss* and weights for the classes, as a solution for the class imbalance in the dataset.

$$L_{CE} = -\sum_{c=0}^{C} w_c y_c \log(\hat{y}_c), \tag{3}$$

where w_c denotes the weight for the class c from the C number of classes, y_c is the true label for class c and $\hat{y}_c$ is the prediction for the same class. We also compare this solution with the proposed in [39] called *Recall Loss*. This loss is computed as:

$$L_{Recall} = -\sum_{c=0}^{C} \frac{FN_c}{FN_c + TP_c} N_c \log(P^c), \tag{4}$$

using the same notation as in the original paper where FN_c are false negatives of class c, TP_c the true positives, N_c the count of c class pixels and P^c is the geometric mean confidence of the class.

For the depth estimation loss L_{Dep} we use the *Adaptive Reverse Huber Loss* (eq. 5), defined as:

$$B_c(e) \begin{cases} |e| & |e| \le c \\ \frac{e^2+c^2}{2c} & |e| > c \end{cases}, \tag{5}$$

where $e = Prediction - GroundTruth$ and c is defined as the 20% of the maximum absolute error for each training batch.

Following the same idea as [27], we also define the loss function as the sum of the *Adaptive Reverse Huber Loss* on the depth map as well as the gradients (approximated as Sobel Filters). The final L_{Dep} is computed as:

$$L_{Dep} = B_{c_1}(e) + B_{c_2}(\nabla_x) + B_{c_2}(\nabla_y), \tag{6}$$

where e defines the absolute depth error between the prediction and ground truth, ∇_x, ∇_y define absolute error between the x, y gradients of the prediction and ground truth respectively, c_1 is the threshold in eq. 5 for the absolute depth map and c_2 is the threshold in eq. 5 for the gradients.

We add another two losses to help in the joint training process. The first one is the margin loss, in order to force the depth estimation branch to fill the depth range between the closest and farthest pixels. We compute the mean square difference between the minimum and maximum values of prediction and ground truth in each batch as:

$$L_{mar} = \frac{\left(y_{gt}^{max} - y_{pred}^{max}\right)^2 + \left(y_{gt}^{min} - y_{pred}^{min}\right)^2}{2}, \tag{7}$$

where y_{gt}^{max}, y_{pred}^{max}, y_{gt}^{min} and y_{pred}^{min} are the maximum and minimum values of the ground truth and predicted depth maps respectively. The second one is an object oriented loss L_{obj}. This loss helps the network to better define the objects boundaries as well as create the depth discontinuities that appear in these boundaries. To compute the loss, we first compute per-class depth maps from the network prediction and ground truth. Then we compute the mean of the *mean squared error* (MSE) of each class depth map to obtain the final L_{obj} as:

$$L_{obj} = \frac{1}{C}\sum_{i=0}^{C}({y_{gt}}^i - {y_{pred}}^i)^2, \tag{8}$$

where C is the number of classes and ${y_{gt}}^i, {y_{pred}}^i$ are the ground truth and predicted class depth maps for the class i respectively.

Our final training loss is the combination of the previous losses. This joint loss function is computed as:

$$L_{total} = \alpha_1 \cdot L_{Seg} + \alpha_2 \cdot L_{Dep} + \alpha_3 \cdot L_{mar} + \alpha_4 \cdot L_{obj}, \tag{9}$$

where α_i are hyper-parameters to weight the relevance of each individual loss in the final joint loss function. At first, we empirically set these hyper-parameters to $\alpha = [8.0, 12.0, 0.001, 4.0]$. After a first training, we make a Bayesian optimization with the tool *Adaptive Experimentation Platform*[1] to optimize these hyper-parameters to obtain the best validation metric for our network.

IV. Experiments on Oriented Panoramas

In this section, we present several ablation studies to evaluate and validate the different design decisions made for our network. We also compare our proposal with state-of-the-art methods that provide depth, semantic segmentation or both tasks. But first, we present our proposal for panoramic image stitching for the Matterport3D dataset.

[1]github.com/facebook/Ax

Fig. 4: Comparison of panoramic stitching. On the left, results of the tool provided by *Joanneum research*. On the right, results of our proposal. Notice the zoom detail where our proposal provides a seamless panorama.

A. Panoramic Stitching of Matterport3DS Dataset

The Matterport3D dataset [11] is a great collection of perspective images taken in indoor environments with an array of cameras to cover 360 degrees field of view. The dataset provides the color images and depth scans as perspective images. Furthermore, they also provide a 3D reconstruction of the different environments as 3D meshes with semantic information in the nodes of the meshes. In order to obtain semantic and depth information in the equirectangular projection, we generate the panoramic images from the raw data provided in the original dataset. We generate the same number of panoramic images as locations are in the dataset. We keep the same train-test split of the original work, providing a multi-modal comparison for future works, i.e. the possibility to compare perspective *vs* panoramic images in the same environments.

To obtain the color panoramas, we blend the perspective images in the equirectangular projection using Gaussian Pyramids [40], reducing the artefacts in the intersection of the perspective images. The blending also takes into account the brightness of the scene, adapting the brightness of the perspective images to avoid hard changes in the illumination conditions in consecutive or overlapping images. Even if the implementation is similar to tools previously proposed[2], we observe great differences in the resulting panorama, as can be seen in Fig. 4 where the intersection of the perspective images is clearly visible. Depth maps are built blending the perspective scans into the equirectangular projection and computing the average depth in overlapping areas. Also, depth measures are corrected with the extrinsic calibration of the array of cameras to compute the distance from focal center of each panorama.

The main novelty that we present are the semantic segmentation maps. These maps are retrieved from the 3D meshes of each environment, where the semantic information is available in the raw data. To retrieve and project this semantic information, we build a virtual environment with the 3D mesh and fill each triangle of the mesh with the label of the nodes.

[2]The proposed tool by *Joanneum Research* is available at https://github.com/atlantis-ar/matterport_utils

TABLE I: Ablation study. Loss functions comparison evaluated on the validation set.

Network	MRE ↓	δ^1 ↑	mIoU ↑	mAcc ↑
F-Std-CE-X	0.0589	0.9059	62.40	**84.47**
F-Std-RL-X	**0.0545**	**0.9172**	**62.69**	79.59
F-Equi-CE-X	0.0626	0.9009	**62.26**	**84.14**
F-Equi-RL-X	**0.0576**	**0.9145**	**62.26**	78.96

TABLE II: Ablation study. Bayesian optimization evaluated on the validation set.

Network	MRE ↓	δ^1 ↑	mIoU ↑	mAcc ↑
F-Std-CE-B	0.0412	0.9404	71.12	**89.76**
F-Std-RL-B	**0.0408**	**0.9409**	**75.61**	88.73
F-Equi-CE-B	**0.0461**	**0.9378**	71.45	**89.15**
F-Equi-RL-B	0.0510	0.9258	**72.56**	86.42

TABLE III: Ablation study. Convolution influence evaluated on the validation set.

Network	MRE ↓	δ^1 ↑	mIoU ↑	mAcc ↑
F-Std-CE-X	**0.0589**	**0.9059**	**62.40**	**84.47**
X-Std-CE-X	0.0629	0.8998	62.27	84.29
F-Equi-CE-X	**0.0626**	**0.9009**	**62.26**	**84.14**
X-Equi-CE-X	0.0666	0.8862	58.43	83.13

Three cases can appear: if the three nodes have the same label, the triangle takes that semantic label; if one node has a different label than the other two, the triangle takes the semantic label of the two nodes; if each node has a different label, the triangle takes no semantic label, that means, that triangle get the "*unknown*" label. Once the virtual environment is built, the image acquisition follows the pipeline described in OmniSCV [41]. We obtain 6 perspective images (a.k.a. cube map) in the same locations as the color images and project these images into the equirectangular projection. We use the extrinsic calibration of the whole array of cameras in each position to correctly align the semantic segmentation maps with the color and depth panoramas obtained from the raw acquisitions.

B. Ablation study

In this section we evaluate different training methods on several network configurations to obtain the network that better generalizes to unknown environments. In [35] we presented ablation studies regarding the joint training and the proposed loss functions with fixed hyper-parameters. Here, we extend these experiments and, for a better understanding, we split the results obtained in 3 experiments. In all the experiments we compare our proposal with standard convolutions and with *EquiConvs*. We show two metrics for each task: Mean Relative Error (*MRE*) and an inlier ratio (δ^1) for depth estimation and the mean Intersection over Union (*mIoU*) and mean Accuracy (*mAcc*) for semantic segmentation. Training is performed in one GPU NVIDIA GeForce RTX3090-Ti with an initial learning rate of 1e-5. We use the first folder split from Stanford2D3DS [42] for training, leaving Area 5 only for test. We make an inner split in the train set (all areas except Area 5) of $80\% - 20\%$ as train-validation sets. We use the validation set for comparison in this ablation study and finally we use the test set to select the network that better generalizes to unknown environments.

We name each network with a four-attribute code where: the first attribute describes if we use an spectral block (**F**) or if we use a convolutional layer instead the Fourier Block (**X**); the second attribute describes if we use standard convolutions (**Std**) or Equiconvs (**Equi**); the third attribute describes the semantic segmentation loss used during training as (**CE**) for the Cross Entropy Loss or (**RL**) for the Recall Loss; and the last attribute describes if we use the Bayesian optimization (**B**) for the tune of hyper-parameters or if we don't (**X**). E.g. the code **F-Std-CE-X** describes a network with FFC and Standard convolutions, trained with the Cross Entropy loss and with hand-set hyper-parameters.

Loss function. In this experiment we compare the performance of the network with the two proposed loss functions for semantic segmentation: *Cross Entropy* and *Recall Loss*. The training of these networks has been done from scratch and with the hyper-parameters set by hand, that is, $\alpha = [8.0, 12.0, 0.001, 4.0]$. From the results shown in in Table I we infer that the Cross Entropy and Recall Loss have similar performance, being better the depth estimation with the Recall loss and better the semantic segmentation with the Cross Entropy (taking into account both semantic metrics). This behaviour may be due to the balance between both branches of the network, which may fight each other for a better performance.

Hyper-parameter optimization. After the first training of the networks, we perform a Bayesian optimization in the hyper-parameters in order to improve their performance. The evaluation results are shown in Table II. When tuning the hyper-parameters with the Bayesian optimization, the metrics obtained during training increase significantly. This can mean two things: the network generalizes better or the fine tune over fits to the validation set.

Convolutions. Additionally, we perform another experiment to evaluate the influence of each type of convolution in the final performance of our network. In this experiment we compare the presented architecture but changing the *Spectral block* from the *Fourier Block* by a convolutional layer, standard or equirectangular. The results shown in Table III strengthen our proposal, the use of convolutions in the frequency domain help the network to better understand the scene.

Performance in unknown environments. Finally, we evaluate the generalization to unknown environments of the several training configurations proposed using the test split of Stanford2D3DS (Area 5 of the dataset). From this evaluation we have selected the two best networks for the state of the art comparison: we take the best network with standard convolutions and the best with EquiConvs. The results are shown in Table IV, where we can see that the Bayesian optimization does not provide a great improvement in the performance of the network, which may lead that this fine tuning over fits the network to the validation set. Nevertheless, these networks are the ones that provide the best performance on gravity-oriented panoramas.

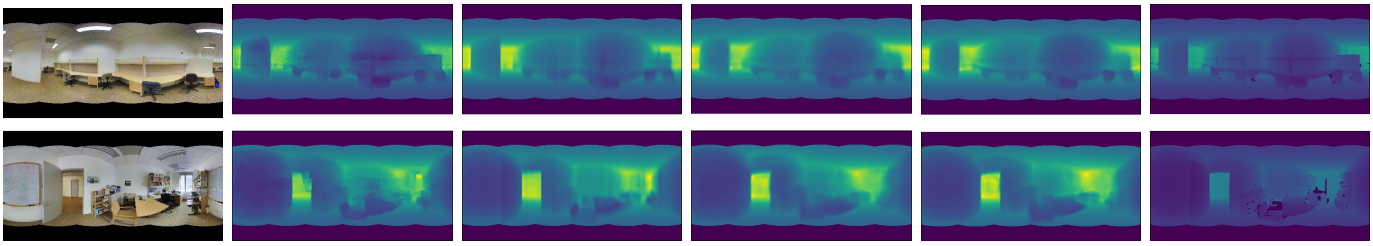


Fig. 5: Qualitative comparison of state-of-the-art methods for depth estimation on gravity-oriented panoramas in the Stanford2D3DS data-set [42].

TABLE IV: Ablation Study. Performance in unknown environments. We evaluate different training methods and proposed architectures on the test set. Selected networks for SOTA comparison are in bold.

Network	MRE ↓	δ^1 ↑	mIoU ↑	mAcc ↑
F-Std-CE-X	0.0961	0.8313	44.85	64.45
F-Std-RL-X	0.0888	0.8416	42.61	58.63
F-Std-CE-B	0.0968	0.8481	45.64	63.76
F-Std-RL-B	0.0964	0.8483	46.31	58.79
X-Std-CE-X	0.0979	0.8147	44.31	64.01
F-Equi-CE-X	0.0993	0.8270	43.29	63.27
F-Equi-RL-X	0.0945	0.8414	42.01	56.31
F-Equi-CE-B	0.1010	0.8533	45.32	62.24
F-Equi-RL-B	0.1046	0.8357	44.82	56.56
X-Equi-CE-X	0.0999	0.7927	42.35	62.90

C. State of the Art Comparison

From the different configurations of our proposal (see section IV-B), we select one that uses standard convolutions (F-Std-CE-B) and other with EquiConvs (F-Equi-CE-B) in parallel with the FFC for the comparison with state-of-the-art methods. The metrics used for the depth estimation task are the standard metrics introduced by [26]. We use the Mean Relative Error (*MRE*), Mean Absolute Error (*MAE*), Mean Square Error of linear (*MSE*) and logarithmic (*MSElog*) measures, and three relative accuracy measures defined as the fraction of pixels where the relative error is within a threshold of 1.25^n for $n = 1,2,3$ ($\delta^1, \delta^2, \delta^3$). On the other hand, for the semantic segmentation task, we use standard metrics as the mean Intersection over Union (*mIoU*), computed as the average *IoU* for each class except the *Unknown* class; and the mean Accuracy (*mAcc*), computed as the average accuracy for each class except the *Unknown* class. The experiments have been conducted using two datasets of panoramic images: Stanford2D3DS [42] and Matterport3D [11] with our presented stitching.

We compare our proposal with task-specific state-of-the-art methods for semantic segmentation task and depth estimation task. We evaluate methods with different architectures, from convolutional-based to attention-based. The methods evaluated are those where network and weights are publicly available and can be tested with the same dataset and metrics.

For the depth estimation task, we compare our work in the Stanford2D3DS dataset [42], following the first folder split, with HohoNet [19] and OmniFusion [29]. We also compare our proposal in the Matterport3D dataset [11] against HohoNet [19] and SliceNet [27].

For the semantic segmentation task, we compare our network with HohoNet[19] and Trans4PASS[21]. We evaluate on the Stanford2D3DS dataset, following the first folder split. We also provide quantitative results of our network in the Panoramic Matterport3DS dataset, with the same labelling as the Stanford2D3DS dataset. Results of other networks are not available since there is no other work that evaluates semantic segmentation on the Matterport3DS dataset with equirectangular panoramas.

Notice that HohoNet appears in both comparisons. This is the only existing method that provides both semantic segmentation and depth estimation from a similar network architecture with public code for evaluation. However, HohoNet uses ground truth depth as well as RGB data as input for semantic segmentation. For a fair comparison, we use the depth output of HohoNet as input for the semantic segmentation task (*this configuration is different from their original experiments*).

The quantitative results of the evaluation of depth estimation and semantic segmentation are presented in Table V and Table VI respectively. We also present a qualitative comparison for depth estimation and semantic segmentation on Figure 5 and Figure 6 respectively.

The experiments show that our method has better performance on semantic segmentation and similar on depth estimation than the only other evaluated method that provides both tasks [19]. In the comparison with task-specific methods, we observe that we are close in performance in each task, but we obtain more information with the same network and at the same time. In particular, in the Stanford2D3DS dataset, we have similar performance than the other two methods in depth estimation and outperform Hohonet by a margin in semantic segmentation, with the same network and weights. The strength of our proposal is presented in the next section where we evaluate these architectures in rotated panoramas. Additionally, the qualitative results of the transformer-based approach shows little inconsistencies on depth estimation maps, being visible the image patches, while convolutional methods provide smoother maps. However, the qualitative comparison on semantic segmentation shows that all methods provide smooth results.

V. Experiments on Rotated Panoramas

Indoor datasets of panoramic images are often aligned with the gravity direction. This means that the distribution of the

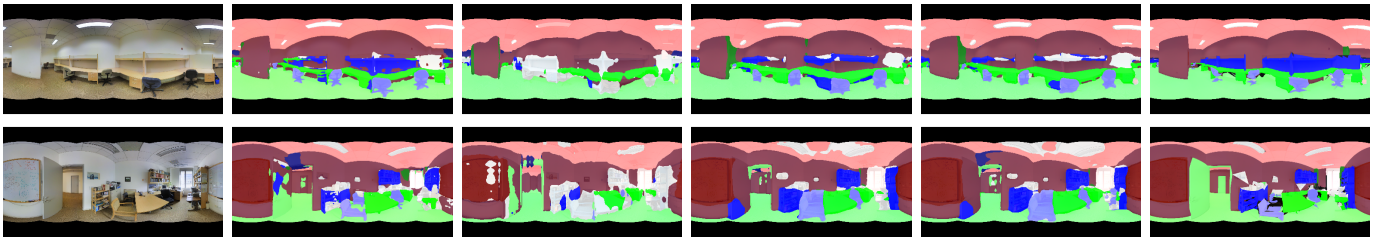


Fig. 6: Qualitative comparison of state-of-the-art methods for semantic segmentation on gravity-oriented panoramas in the Stanford2D3DS data-set [42].

TABLE V: Quantitative comparison for Depth Estimation on gravity-oriented panoramas. In **bold** are the two best metrics.

Dataset	Network	MRE ↓	MAE ↓	MSE ↓	MSElog ↓	δ^1 ↑	δ^2 ↑	δ^3 ↑
Stanford2D3DS[42]	HohoNet[19]	**0.0812**	**0.1162**	0.3137	**0.0367**	**0.8838**	**0.9674**	**0.9880**
	OmniFusion[29]	**0.0798**	0.1566	**0.1443**	0.0480	**0.8619**	**0.9654**	0.9771
	F-Std-CE-B	0.0968	0.1320	0.2738	**0.0430**	0.8481	0.9589	**0.9863**
	F-Equi-CE-B	0.1010	**0.1305**	**0.2638**	0.0439	0.8533	0.9579	0.9848
Matterport3D[11]	HohoNet[19]	0.8847	0.4170	**0.5610**	**0.0576**	**0.8409**	**0.9483**	**0.9781**
	SliceNet[27]	0.8434	0.4398	**0.4836**	0.1036	**0.7974**	**0.9231**	0.9671
	F-Std-CE-B	**0.7220**	**0.3815**	0.7036	**0.0782**	0.7517	0.9201	**0.9695**
	F-Equi-CE-B	**0.7022**	**0.3883**	0.6635	0.0846	0.7366	0.9125	0.9658

TABLE VI: Quantitative comparison for Semantic Segmentation on gravity-oriented panoramas. Greater metrics is better. In **bold** are the two best metrics.

Dataset	Network	mIoU	mAcc
Stanford 2D3DS[42]	HohoNet[19]	35.6	45.7
	Trans4PASS-T[21]	**47.7**	58.2
	Trans4PASS-S[21]	**52.3**	**62.7**
	F-Std-CE-B	45.6	**63.8**
	F-Equi-CE-B	45.3	62.2
Panoramic Matterport3DS	HohoNet[19]	-	-
	Trans4PASS-T[21]	-	-
	Trans4PASS-S[21]	-	-
	F-Std-CE-B	**35.4**	**57.4**
	F-Equi-CE-B	**37.3**	**52.5**

environment is almost always the same: floor at the bottom of the image, ceiling at the top and walls in the middle. However, for indoor navigation systems, this particular orientation could not be completely fulfilled (i.e. when flying a drone indoors, bi-pedal or quad-pedal robots or taking the camera by hand), therefore this distribution can change. In this section, we evaluate the robustness of different networks to non-gravity-oriented panoramas. To obtain these images, we take the images from the Stanford2D3DS dataset [42] and rotate these panoramas around a horizontal axis.

A. Ablation Study

We evaluate how the proposed convolutions and training methods behave under rotated panoramas. All networks have been trained on gravity oriented panoramas in the Stanford dataset. The evaluation is made in the test set (*Area 5*) rotating the panoramas around an axis in the horizon plane and a fixed angle. Results of this experiment are shown in Figure 7. Additionally, we present qualitative results of success and failure cases of rotated panoramas in Figure 9.

In the hyper-parameters' comparison (see curves with *B* attribute at Figure 7) we observe that, for small rotation angles, networks with the Bayesian optimization provide the best performance. However, as we increase the angle of rotation, these networks decrease their performance faster than any other, ending as the worst solution. Our intuition is that networks with Bayesian optimized hyper-parameters over fit the validation data (used for the hyper-parameter tuning), generalizing worse on more general environments. In the comparison of convolutions, on depth estimation (see first two graphs on Figure 7) we can observe that, for all the angles, the networks with equirectangular convolutions provide the best performance, *F-Equi-CE-B* at first and *F-Equi-CE-X* for greater angles. However, for semantic segmentation (see last two graphs on Figure 7) this behaviour does not apply and *F-Std-CE-X* prevails as the best option for almost all angles. These results show that equirectangular convolutions may not provide the best performance for gravity-oriented equirectangular panoramas. Nevertheless, we do see a trend where Equiconvs are more robust than standard convolutions on non-gravity-oriented panoramas, presenting less performance decay.

B. State of the Art Comparison.

In this experiment, we compare different state-of-the-art networks and evaluate how well they generalize to non-gravity-oriented panoramas. For that purpose, we test the networks as in the previous experiment. Additionally, we fine tune the networks on non-gravity-oriented panoramas to evaluate their behaviour when they have learned under these circumstances. To differentiate the networks trained on non-gravity-oriented panoramas, we include at the end of their name the attribute "-r". The training is made rotating the panoramas around a random axis in the horizon plane and an angle sampled from an uniform distribution $\mathcal{U}(-10^{\circ}, 10^{\circ})$. The evaluation is made rotating the panoramas around an axis in the horizon plane and a fix angle between 0 and 30 degrees. Results of this experiment are shown in Figure 8.

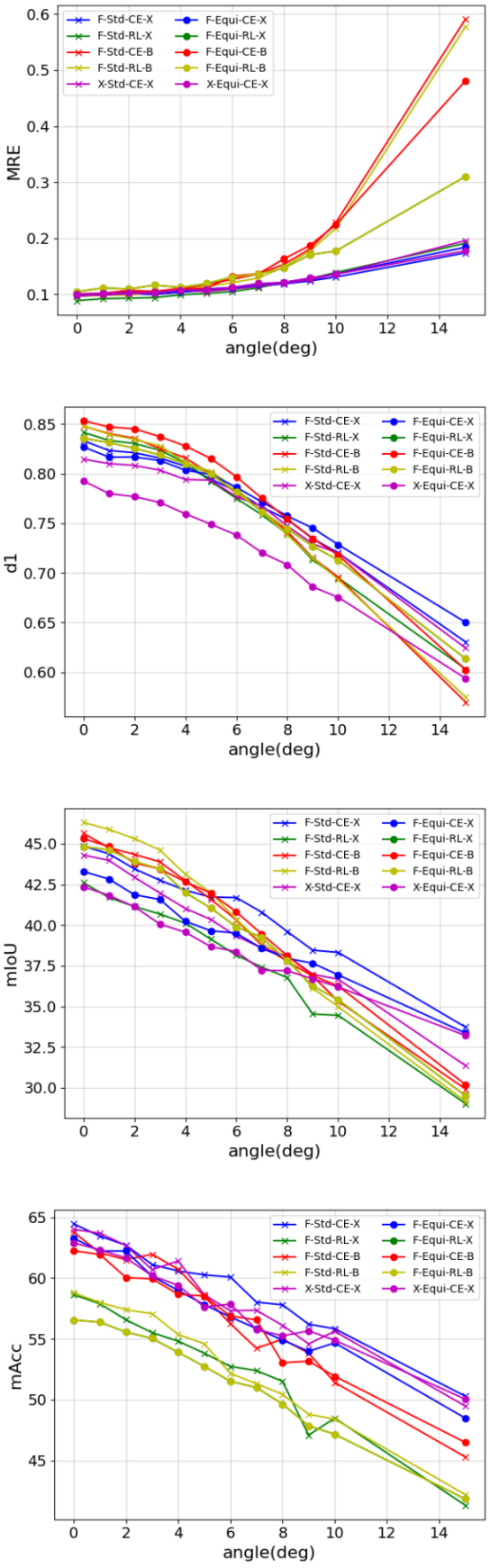


Fig. 7: Ablation Study of the different training and architectures under different rotation angles.

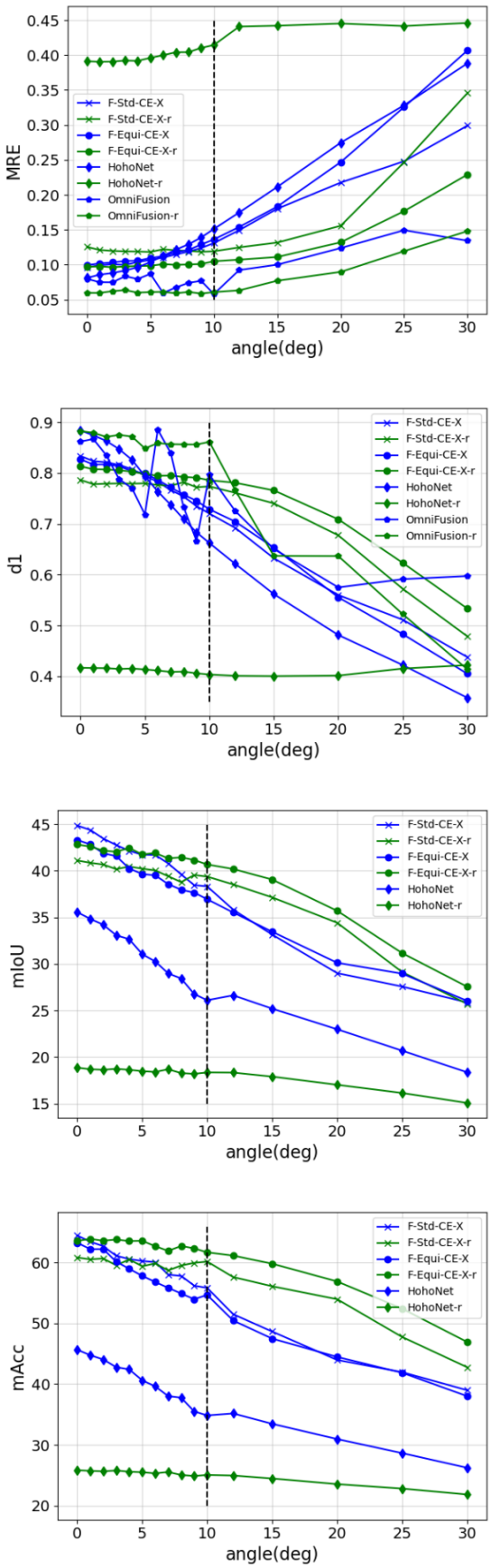


Fig. 8: Comparison of the different state-of-the-art methods under different rotation angles. The methods which name finish as "-r" have been trained on rotated panoramas.

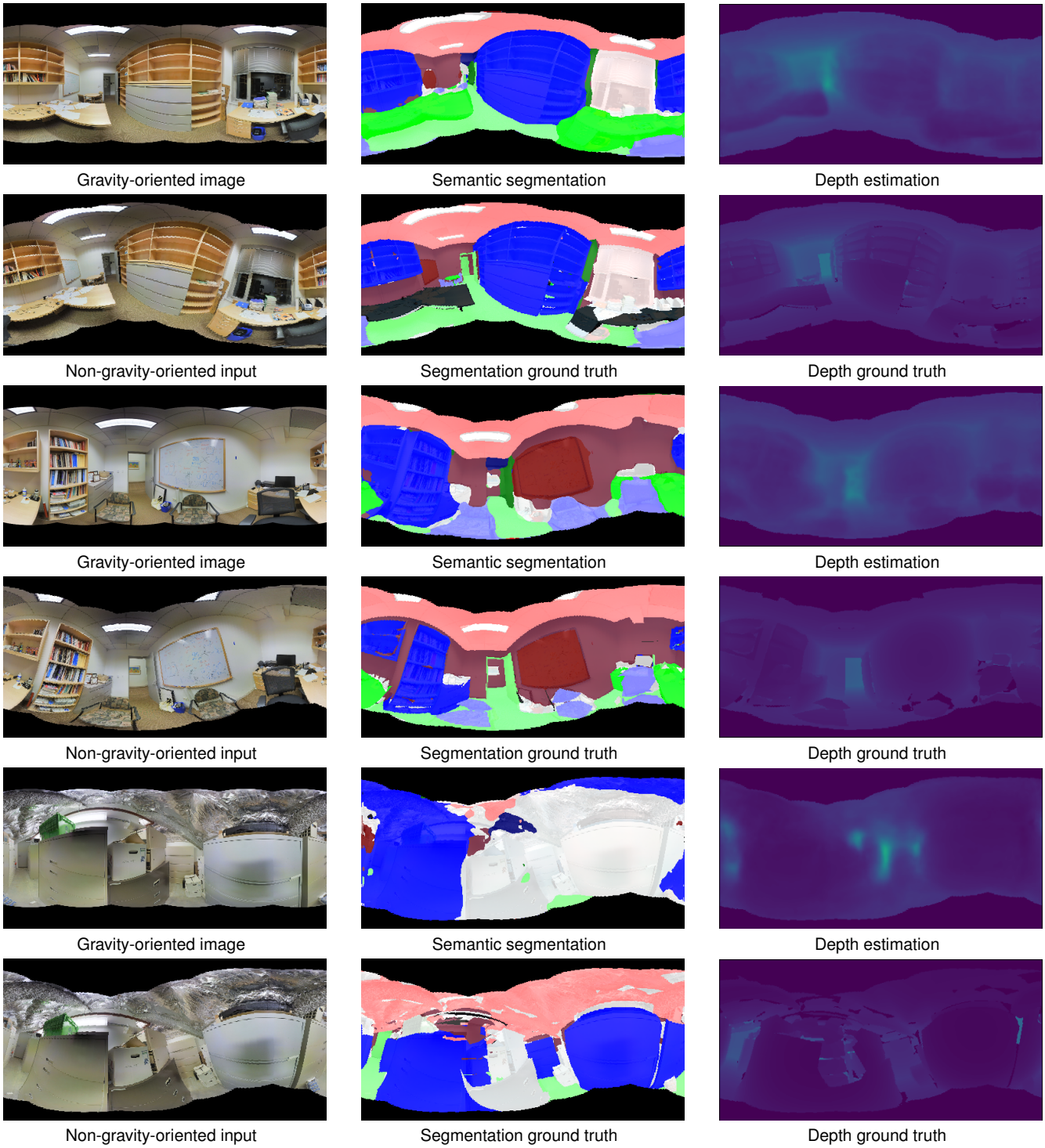


Fig. 9: Qualitative result of non-gravity oriented panoramas rotated 15° around an axis in the $X-Y$ plane. In the first two examples (first 4 rows) we present successful cases while in the last example (last two rows) we present the worst found case. Results are from our network on rotated panoramas, F-Equi-CE-X-r.

From the results presented of our proposal, we observe that the initial performance of the networks trained on gravity-oriented panoramas is a bit better than the networks trained on non-gravity-oriented panoramas. However, as we increase the rotation angle, those trained on non-gravity-oriented panoramas are significantly more robust. We also observe that the network with the Equiconvs performs better that the network with standard convolutions, validating their use on more general conditions.

Analysing the performance of OmniFusion[29], we observe that it has an uneven behaviour while the rotation angle increases. On the other hand, OmniFusion-r has greater per-

formance and a more stable behaviour, being the network with best performance in the training range (i.e. up to 10°). However, it suffers a bigger performance decay for new rotation angles than our proposed method. Our intuition is that the division of the rotated panoramas into patches for the visual transformer makes OmniFusion to struggle at some angles, while does not affect at others. Besides, we believe that OmniFusion-r learns the rotation while training but struggles to generalize outside training data, obtaining even worse results than OmniFusion.

The results of HohoNet[19] for depth estimation show that has a great performance with small angles, but it decays really fast as the rotation angle increases. For the semantic segmentation task, the performance also decreases faster than on other methods. On the other hand, HohoNet-r presents really bad results from the beginning. Even if the performance is more constant than HohoNet, the quantitative results are worse than other methods. We think that the 1D representation of this architecture cannot handle different orientations during training, which leads to worse performances on evaluation.

After these experiments we believe that the 1D representation of the environment of HohoNet and the patch division of OmniFusion does not allow the networks to generalize to different orientations of the panorama outside the training data while the fast Fourier convolutions and EquiConvs are more suited for more general conditions. In this aspect, our proposal is able to better adapt to a more general problem, obtaining more robust and stable results in unknown conditions and outperforming the state of the art in non-gravity-oriented panoramas.

VI. Conclusion

In this paper we present a novel approach to jointly estimate monocular depth and semantic segmentation from panoramas in challenging conditions, such as non-gravity oriented panoramas. We have evaluated different components of common convolutional architectures, observing that the combination of fast Fourier convolutions and Equirectangular convolutions does improve the performance and robustness of our proposal with rotated panoramas, providing a better generalization of the network on more general and challenging conditions.

In the comparison with the state of the art, we observe how our proposal have a more predictable behaviour, achieving the best performance for unknown conditions in both, depth estimation and semantic segmentation against single task neural networks. Besides, our experiments on oriented panoramas show that our proposal has slightly better performance than the sole state-of-the-art method that obtains both tasks, depth estimation and semantic segmentation in equirectangular images. Even though, task specific methods can achieve better performance on each individual task, our proposal is able to obtain both tasks simultaneously and from a single RGB panorama.

Our current approach can only handle the distortion of panoramic images. In a future work, we would like to extend this idea to different imagery, such as fisheye or catadioptric systems. Besides, our neural network is too large to be incorporated in portable devices. For future applications, we would like to reduce the dimensionality and requirements of our proposal to be able to work on wearable devices or autonomous robots. Finally, one limitation of the method can be seen in Figure 9. Our proposal is the one with the best performance in rotated panoramas and unknown rotations, however it still struggles with big rotations. As future work, we will improve our proposal, finding new methods to handle the image rotation.

Acknowledgments

This work was supported by projects PID2021-125209OB-I00 and TED2021-129410B-I00 (MCIN/AEI/10.13039/501100011033 and FEDER/UE and NextGenerationEU/PRTR).

References

[1] V. Arampatzakis, G. Pavlidis, N. Mitianoudis, and N. Papamarkos, "Monocular depth estimation: A thorough review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2396–2414, 2024.

[2] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[3] Z. Xie, X. Yu, X. Gao, K. Li, and S. Shen, "Recent advances in conventional and deep learning-based depth completion: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2022.

[4] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "Orb-slam: a versatile and accurate monocular slam system," *IEEE transactions on robotics*, pp. 1147–1163, 2015.

[5] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 12, pp. 6999–7019, 2022.

[6] M. Heo, J. Lee, K.-R. Kim, H.-U. Kim, and C.-S. Kim, "Monocular depth estimation using whole strip masking and reliability-based refinement," in *European Conference on Computer Vision*. Springer, 2018, pp. 36–51.

[7] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun, "Dense monocular depth estimation in complex dynamic scenes," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2016.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 2961–2969.

[9] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," in *Conference on Neural Information Processing Systems*, 2020, pp. 4479–4488.

[10] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero, "Corners for layout: End-to-end layout recovery from 360 images," *Robotics and Automation Letters*, pp. 1255–1262, 2020.

[11] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," in *International Conference on 3D Vision*. IEEE, 2017, pp. 667–676.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 211–252, 2015.

[13] W. Su and Z. Wang, "Widening residual skipped network for semantic segmentation," *IET Image Processing*, vol. 11, no. 10, pp. 880–887, 2017.

[14] Z. Yang, H. Yu, Y. He, W. Sun, Z.-H. Mao, and A. Mian, "Fully convolutional network-based self-supervised learning for semantic segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 132–142, 2024.

[15] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "Blitznet: A real-time deep network for scene understanding," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 4154–4162.

[16] R. Cong, K. Huang, J. Lei, Y. Zhao, Q. Huang, and S. Kwong, "Multi-projection fusion and refinement network for salient object detection in 360 omnidirectional image," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2023.

[17] M. Eder, M. Shvets, J. Lim, and J.-M. Frahm, "Tangent images for mitigating spherical distortion," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2020, pp. 12 426–12 434.

[18] J. Guerrero-Viu, C. Fernandez-Labrador, C. Demonceaux, and J. J. Guerrero, "What's in my room? object recognition on indoor panoramic images," in *International Conference on Robotics and Automation*. IEEE, 2020, pp. 567–573.

[19] C. Sun, M. Sun, and H.-T. Chen, "Hohonet: 360 indoor holistic understanding with latent horizontal features," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2021, pp. 2573–2582.

[20] K. Yang, X. Hu, and R. Stiefelhagen, "Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild," *IEEE Transactions on Image Processing*, vol. 30, pp. 1866–1881, 2021.

[21] J. Zhang, K. Yang, C. Ma, S. Reiß, K. Peng, and R. Stiefelhagen, "Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2022, pp. 16 917–16 927.

[22] D. Cao Dinh, S. J. Kim, and K. Cho, "Geometric exploitation for indoor panoramic semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 26 355–26 376, 2024.

[23] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, "Cam-convs: Camera-aware multi-scale convolutions for single-view depth," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2019, pp. 11 826–11 835.

[24] J.-H. Lee and C.-S. Kim, "Monocular depth estimation using relative depth maps," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2019, pp. 9729–9738.

[25] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu, "Unidepth: Universal monocular metric depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 10 106–10 116.

[26] N. Zioulis, A. Karakottas, D. Zarpalas, and P. Daras, "Omnidepth: Dense depth estimation for indoors spherical panoramas," in *European Conference on Computer Vision*. Springer, 2018, pp. 448–465.

[27] G. Pintore, M. Agus, E. Almansa, J. Schneider, and E. Gobbetti, "Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2021, pp. 11 536–11 545.

[28] J. Lee, H. Park, B.-U. Lee, and K. Joo, "Hush: Holistic panoramic 3d scene understanding using spherical harmonics," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 16 599–16 608.

[29] Y. Li, Y. Guo, Z. Yan, X. Huang, Y. Duan, and L. Ren, "Omnifusion: 360 monocular depth estimation via geometry-aware fusion," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2022, pp. 2801–2810.

[30] F.-E. Wang, Y.-H. Yeh, M. Sun, W.-C. Chiu, and Y.-H. Tsai, "Bifuse: Monocular 360 depth estimation via bi-projection fusion," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 2020, pp. 462–471.

[31] F.-E. Wang, Y.-H. Yeh, Y.-H. Tsai, W.-C. Chiu, and M. Sun, "Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)*, 2022.

[32] C. Han, Y. Cai, X. Pan, and Z. Wang, "Effective fusion module with dilation convolution for monocular panoramic depth estimate," *IET Image Processing*, vol. 18, no. 4, pp. 1073–1082, 2024.

[33] H. Ai and L. Wang, "Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9926–9935.

[34] J. Bai, H. Qin, S. Lai, J. Guo, and Y. Guo, "Glpanodepth: Global-to-local panoramic depth estimation," *IEEE Transactions on Image Processing*, 2024.

[35] B. Berenguel-Baeta, J. Bermudez-Cameo, and J. J. Guerrero, "Fredsnet: Joint monocular depth and semantic segmentation with fast fourier convolutions from single panoramas," in *International Conference on Robotics and Automation*. IEEE, 2023, pp. 6080–6086.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 770–778.

[37] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the International Conference on Computer Vision*. IEEE, 2017, pp. 764–773.

[38] Z. Zhang, Z. Cui, C. Xu, Z. Jie, X. Li, and J. Yang, "Joint task-recursive learning for semantic segmentation and depth estimation," in *European Conference on Computer Vision*. Springer, 2018, pp. 235–251.

[39] J. Tian, N. C. Mithun, Z. Seymour, H.-P. Chiu, and Z. Kira, "Striking the right balance: Recall loss for semantic segmentation," in *International Conference on Robotics and Automation*. IEEE, 2022.

[40] P. J. Burt and E. H. Adelson, "A multiresolution spline with application to image mosaics," *Transactions on Graphics*, vol. 2, no. 4, pp. 217–236, 1983.

[41] B. Berenguel-Baeta, J. Bermudez-Cameo, and J. J. Guerrero, "Omniscv: An omnidirectional synthetic image generator for computer vision," *Sensors*, 2020.

[42] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *2017, arXiv preprint arXiv:1702.01105*, 2017.