ATLANTIS Technical Evaluation

ATLANTIS Public Report Nr. 9	
Project:	ATLANTIS – AuThoring tooL for indoor Augmented and dimiNished realiTy experIenceS
Website:	http://atlantis-ar.eu
Author(s):	Vasileios Gkitsas, Petros Drakoulis, Antonis Karakottas, Alexandros Doumanoglou (CERTH), Werner Bailer & Georg Thallinger (JOANNEUM RESEARCH), Robert Huemer (ROOMLE)
Publication date:	2022-07-07
Version number:	1.1
Abstract:	This report provides an overview of the technical evaluations complementing the user experiments. It provides details on the datasets and metrics used, and provides results and pointers to published results.

European Commission | This project has received funding from the European Union's Horizon 2020 Innovation Action programme under **grant agreement No 951900**

Introduction

In addition to the user tests (see Report 6) a number of technical evaluations on the level of components or groups of components have been performed. This report provides an overview of the tests that have been performed on the level of components or integrated pipelines, and the datasets and metrics used.

A summary of the tests is provided in Table 1, and the following sections describe tests in more detail or provide pointers to additional material.

Table 1: Summary of the component and integrated tests, and the datasets and metrics used.

Test	Datasets	Metrics
Component tests		
Room layout estimation	Sun360, Stanford2D3D, Structured3D	CE, PE, IoU2D, IoU3D, Junction and Wireframe Accuracies, Depth RMSE, Depth Accuracy
Depth estimation	Matterport3D, Structured3D, GibsonV2, Stanford2D3D	RMSE, RMSLE, 3 levels of relative accuracy, mean absolute relative error, mean squared relative error, precision and recall of depth boundaries, completeness and accuracy of depth edges, surface orientation angular RMSE, surface orientation accuracy at 3 different thresholds, Hausdorff distance on 3D reconstructed meshes, point cloud distance on deprojected point clouds
Instance segmentation	Matterport3D, ScanNet, Structured3D	MAP at specific IoUs, pixelwise mask accuracy, runtime
Scene graph estimation	Structured3D	CE, PE, Box IoU2D, Junction and Wireframe Accuracies, Object Detection MAP, 3D object bounding box RMSE, SupErr, Layout+boxes depth accuracy
Inpainting for DR	Structured3D	MAE, PSNR, SSIM, LPIPS
Scene layout completion	Roomle 10K plans	statistical comparison to ground truth samples
On-device segmentation update	ScanNet	MAP at specific IoUs
Integrated tests		
Instance segmentation + layout + inpainting	Structured3D	MAE, PSNR, SSIM, LPIPS

Room layout estimation

For the purposes of ATLANTIS, we developed a novel cuboid-room sparse layout prediction model which was published under the peer-reviewed *"Single-Shot Cuboids: Geodesics-based End-to-end Manhattan Aligned Layout"*[1], *accompanied by its respective GitHub repository* [2]. At the time of its launch, it achieved state-of-the-art performance, as summarised in the following tables:

[1] https://arxiv.org/pdf/2102.03939.pdf
[2] https://vcl3d.github.io/SingleShotCuboids/

Table 2: Quantitative results on two real domain datasets. SSC is our model. Red denotes the best-performing model and orange the second-best.

Model			PanoContext			Stanford2D3D		
Name	Variant	Parameters ↓	CE ↓	IoU3D ↑	PE ↓	CE ↓	IoU3D ↑	PE ↓
LayoutNet v2	ResNet-34	25.68M	0.63%	85.02%	1.79%	0.71%	84.17%	2.04%
DuLa-Net v2	ResNet-50	57.38M	0.81%	83.77%	2.43%	0.67%	86.6%	2.48%
HorizonNet	ResNet-50	81.57M	0.74%	82.63%	2.17%	0.69%	82.72%	2.27%
SSC	**HG-3**	**6.35M**	**0.63%**	**83.97%**	**1.78%**	**0.51%**	**87.80%**	**1.62%**

Table 3: Quantitative results on the synthetic Structured3D dataset. SSC is our model. Red denotes the best-performing model.

Model	Variant	CE ↓	IoU2D ↑	IoU3D ↑	PE ↓	J5 ↑	J10 ↑	J15 ↑	W5 ↑	W10 ↑	W15 ↑
Horizon Net ResNet-50	Post-processed	0.75%	93.49%	91.82%	1.46%	78.61%	91.64%	95.75%	56.67%	78.22%	87.57%
SSC HG-3	**w/ Homography (joint)**	**0.4%**	**94.27%**	**92.33%**	**1.26%**	**75.35%**	**90.9%**	**95.74%**	**48.16%**	**74.35%**	**85.68%**

Table 4: Cross-validation results on the Kujiale dataset using the Structured3D trained model. SSC is our model. Red denotes the best-performing model.

Model	Variant	CE ↓	IoU2D ↑	IoU3D ↑	PE ↓	J5 ↑	J10 ↑	J15 ↑	W5 ↑	W10 ↑	W15 ↑
Horizon Net ResNet-50	Post-processed	1.04%	90.97%	88.96%	1.82%	71.18%	86.59%	92.95%	45.45%	69.67%	81.39%
SSC HG-3	**w/ Homography (joint)**	**0.42%**	**93.37%**	**91.21%**	**1.38%**	**71.82%**	**87.36%**	**94.86%**	**44.12%**	**70.73%**	**82.06%**

Depth estimation

To ensure state-of-the-art performance from the depth estimation component we conducted a thorough ablation study, a peer-reviewed benchmark, involving many popular architectures and supervision schemes, suitable for the task of monocular 360 depth estimation. Our work is published under *"Pano3D: A*

holistic benchmark and a solid baseline for 360 depth estimation"[3] and is accompanied by a publicly available GitHub repository accessible at [4]. A summary of results found in this work, justifying our model choice, is presented below:

Table 5: Models' raw depth estimation performance. Unet^{vnl} is the model chosen to be served by the AI-Service. Red denotes the best-performing model and orange the second-best.

Model	Error ↓				Accuracy ↑				
	wRMSE	wRMSLE	wAbsRel	wSqRel	δ $ico^6_{1.05}$	δ $ico^6_{1.1}$	δ $ico^6_{1.25}$	δ $ico^6_{1.25^2}$	δ $ico^6_{1.25^3}$
Pnas^{comb}	0.5367	0.0811	0.1259	0.1153	36.44%	60.52%	86.80%	95.83%	98.11 %
Unet^{vnl}	**0.4520**	**0.1300**	**0.1147**	**0.0811**	**36.68%**	**60.59%**	**88.31%**	**96.96%**	**98.73 %**
DenseNet^{comb}	0.5209	0.1982	0.1209	0.1013	35.97%	60.41%	87.02%	95.96%	98.09 %
ResNet^{comb}	0.5294	0.1365	0.1374	0.1127	32.03%	55.31%	84.74%	95.81%	98.21 %
ResNet$^{comb}_{skip}$	0.4788	0.0927	0.1166	0.0893	36.20%	60.64%	87.99%	96.62%	98.49 %

Table 6: Models' boundary-preservation performance. Unet^{vnl} is the model chosen to be served by the AI-Service. Red denotes the best-performing model and orange the second-best.

Model	Error ↓		Accuracy ↑					
	dbe^{acc}	dbe^{comp}	$prec_{0.25}$	$prec_{0.5}$	$prec_1$	$rec_{0.25}$	$rec_{0.5}$	rec_1
Pnas^{comb}	2.5119	5.3501	39.83%	31.59%	27.01%	23.53%	14.42%	10.98%
Unet^{vnl}	**1.2699**	**3.8876**	**58.97%**	**57.54%**	**51.85%**	**43.96%**	**36.69%**	**28.59%**
DenseNet^{comb}	2.0628	5.0977	47.16%	40.77%	35.20%	26.09%	16.87%	12.21%
ResNet^{comb}	2.2393	5.3796	44.10%	36.70%	27.44%	22.91%	12.23%	7.20%
ResNet$^{comb}_{skip}$	1.4883	4.5346	57.34%	54.11%	47.57%	33.99%	24.30%	16.37%

[3] https://openaccess.thecvf.com/content/CVPR2021W/OmniCV/papers/Albanis_Pano3D_A_Holistic_Benchmark_and_a_Solid_Baseline_for_360deg_CVPRW_2021_paper.pdf

[4] https://vcl3d.github.io/Pano3D/

Table 7: Models' depth-smoothness performance. Unet^{vnl} is the model chosen to be served by the AI-Service. Red denotes the best-performing model and orange the second-best.

Model	Error ↓	Accuracy ↑		
	RMSE°	$\alpha_{11.25°}$	$\alpha_{22.5°}$	$\alpha_{30°}$
Pnas^{comb}	15.26	67.73%	77.99%	81.67%
Unet^{vnl}	**16.02**	**61.80%**	**76.58%**	**81.70%**
DenseNet^{comb}	15.98	64.58%	76.86%	81.20%
ResNet^{comb}	16.63	63.09%	75.70%	80.20%
ResNet$^{comb}_{skip}$	15.27	64.18%	77.57%	82.27%

Inpainting for DR

For the initial results, please refer to the corresponding publication "*PanoDR: Spherical Panorama Diminished Reality for Indoor Scenes*". Here, we present a summary of results on the Structured3D dataset, the only available dataset providing full and empty configuration and thus capable of evaluating diminished reality. We showcase the effectiveness of our method (PanoDR and Improved PanoDR) that surpass state-of-the-art methods both quantitatively and qualitatively. In the table below, a summary of quantitative results can be found, while for qualitative comparisons please refer to the aforementioned publication.

Table 8: Evaluation of Inpainting models in Diminished Reality. The best-performing model is highlighted in red, while the second-best model is highlighted in orange.

Method	LPIPS ↓	PSNR↑	SSIM↑	mIoU↑
RFR[5]	0.0510	31.0114	0.9528	0.8583
PICNet[6]	0.0533	32.3072	0.9557	0.8502
Ours (PanoDR)[7]	0.0398	33.6611	0.9620	0.8768
Ours (Improved PanoDR)[8]	**0.0320**	**33.6576**	**0.9624**	**0.8789**

[5] Li, J., Wang, N., Zhang, L., Du, B., & Tao, D. (2020). Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7760-7768).

[6] Zheng, C., Cham, T. J., & Cai, J. (2019). Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1438-1447).

[7] https://openaccess.thecvf.com/content/CVPR2021W/OmniCV/html/Gkitsas_PanoDR_Spherical_Panorama_Diminished_Reality_for_Indoor_Scenes_CVPRW_2021_paper.html

[8] https://arxiv.org/abs/2112.05396

Rather than using a pre-trained network for layout estimation, a generative model is used to learn the scene layout feature space. So, both issues in the training process were avoided, and performance was boosted both qualitatively and quantitatively. The publication entitled "Towards Full-to-Empty Room Generation with Structure-Aware Feature Encoding and Soft Semantic Region-Adaptive Normalisation" provides more details. The quantitative evaluation of our model (Improved PanoDR) clearly illustrates the performance boost. Thus, we deploy our Inpainting method as a service to facilitate AR with Diminished Reality.

Instance segmentation

We primarily report results on the Scannet benchmark, as we are interested in the performance of the methods on real rather than synthetic data. It has to be noted that there is no benchmark dataset for indoor instance segmentation on panoramic images. Scannet runs an official benchmark, thus the test set is not publicly available. It is possible to evaluate results on the dataset, but only a limited number of submissions is possible. The official benchmark includes 18 objects classes from the NYU40 set (not using the wall and floor which is otherwise in the commonly used subset of 20 classes). For this reason other results are reported on the Scannet validation set.

Table 9 reports results on the Scannet benchmark set, comparing it to UniDet_RVC, the best reported 2D instance segmentation result on the Scannet benchmark[9]. The results of the method using the COCO backbone are close to the best known results on that dataset (with an AP of 0.205). The COCO backbone improves the results over a model trained from scratch on Scannet. It is also worth noting, although the wall and floor classes are not evaluated in the benchmark, the model including them outperforms the model with just 18 classes. For a number of classes, our model provides the best results on the Scannet benchmark set.

Table 9: Results on the Scannet benchmark.

Metric	**AP**	**AP**	**AP**	**AP**	**AP50**	**AP50**	**AP50**	**AP50**
Method	**UniDet_ RVC**	**SOLOv2 18 classes, COCO back-bone**	**SOLOv2 20 classes**	**SOLOv2 18 classes**	**UniDet_ RVC**	**SOLOv2 18 classes, COCO back-bone**	**SOLOv2 20 classes**	**SOLOv2 18 classes**
mean AP	0.205	0.195	0.175	0.166	0.358	0.321	0.298	0.297

We trained a segmentation model for additional 14 classes mined from the Scannet dataset, which were found relevant for an indoor planning application (board, nightstand, ceiling, pillow, door frame, radiator, garbage bin, shower wall, lamp, stool, microwave, telephone, mirror, whiteboard). The dataset was split into a training and validation set, and the results provided here are those on the validation set. The mean AP for these new classes is 0.149.

[9] http://kaldir.vc.in.tum.de/scannet_benchmark/

Scene layout completion

Objective evaluation using Fréchet inception distance (FID) was utilised. This metric measures statistical similarity between ground truth and generated layouts. Assessing generative approaches is challenging, as one aim is also to obtain diversity. Thus strict comparison against a ground truth will only measure the ability of the model to mimic the ground truth, but not to generalise and produce new variants. It has to be noted that the FID scores are not directly correlated to human assessment of the proposed layouts. We thus provide a comparison of the distribution obtained from calculating the metric on a set of human created layout vs. a set of generated layouts. We use a dataset of 10,000 user created floor plans provided by Roomle for this experiment. The data has been filtered to eliminate incompletely furnished rooms.

Figure 1 shows FID scores obtained from validation (real) and generated samples. It is obvious that the validation samples are narrowly distributed (with few outliers), as they are drawn from the same distribution as the samples on which the model used for determining the FID scores has been trained. The generated layouts have a significantly higher FID, meaning they are clearly distinguishable from real samples. However, as indicated by the lower whisker of the plot, we see it comes close to the range of real samples, which means that the generator is able to create a number of fairly realistic looking samples.

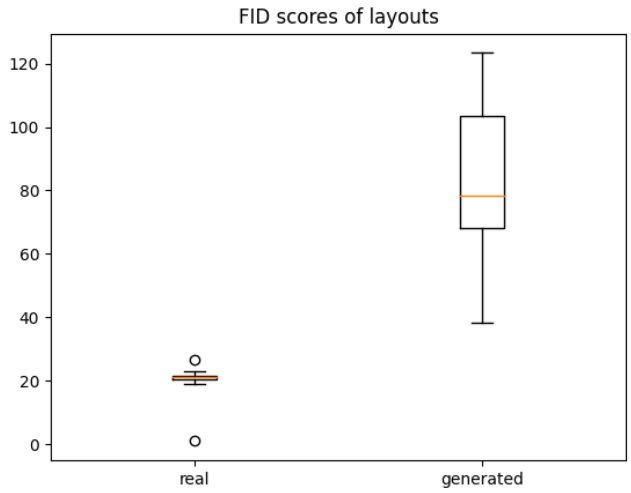

Figure 1: FID scores for real and generated layouts.

On-device segmentation update

Yolact++ is used as the segmentation algorithm on the mobile device, employing a ResNet-50 backbone. We use both models trained on the 80 COCO classes (measured on the test-dev split) and 20 NYU40 classes (using the Scannet dataset, trained for 45 epochs, measured on the validation split). The results are reported in Table 10.

Table 10: Average precision of Yolact++ segmentation with models trained on MS COCO and Scannet (mask AP).

	AP	AP @0.50 IoU	AP @0.75 IoU
MS COCO	34.1	53.3	36.2
Scannet	22.7	37.9	23.2

Instance segmentation + layout + inpainting

Apart from the per component testing described in the subsections above, the ATLANTIS AI services need to interoperate as a system to drive the DR-enhanced AR experiences of users. Towards that end, the systematic evaluation of the entire pipeline needs to be conducted to identify potential integration - and other - issues early on. ATLANTIS developed a service orchestration tool to facilitate such validation experiments. Specifically to support the first workflow, namely the replacement of an object using AR, a validation experiment was conducted. This also included an out-of-the-box super-resolution AI service[10], as well as the compositing of objects from Roomle's catalogue. The goal was to conduct subjective tests to assess the efficacy of DR. On a 1-5 scale (1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent), participants rated the inpainting quality on average with 2.8, and the quality of inserted AR objects on average with 3.6.

More information

Visit https://atlantis-ar.eu or follow us on Twitter @AtlantisAR.

[10] Zhao, H., Kong, X., He, J., Qiao, Y. and Dong, C., 2020, August. Efficient image super-resolution using pixel attention. In *European Conference on Computer Vision* (pp. 56-72).

