Monocular image Depth estimation, using virtual datasets and embedding the semantic information

Christian Sciuto
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
christian.sciuto@epfl.ch

Abstract

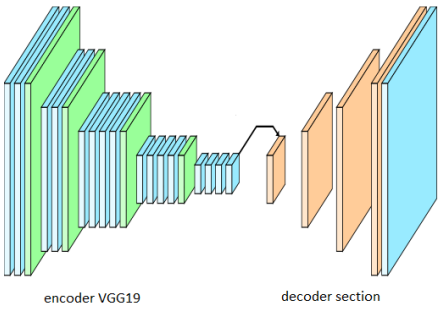
Depth estimation is a fundamental tool for understanding the scene represented in an image, it underpins more complex computer vision tasks and it is of critical importance in various fields such as robotics and self-guidance. The extraction of depth information from individual RGB images is of particular interest, and several works have tried to address this problem through deep learning techniques. In this study we want to investigate the possibility of training a model using only synthetic data to predict a depth map given a single image, i. e. video sequences of a virtual world reproduced by an engine. At the same time we will try to understand the impact on the performance of our network in the situation where it is optimized simultaneously on two related tasks, in our case depth estimation and semantic segmentation of a scene. We will compare two networks to verify our ideas, testing on both virtual and real data.

1 Introduction

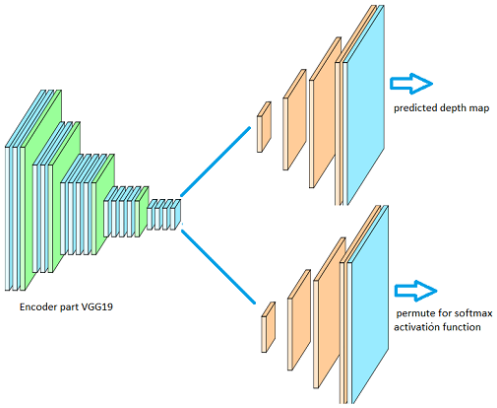
In the field of computer vision, depth estimation is a fundamental task for understanding the scene represented in an image. Being a low level abstraction task, depth estimation is used as a basic component in different areas such as robotics and for autonomous vehicles. In addition, it is now common knowledge that the usage of depth information of a scene offers a huge benefit to other computer vision activities such as semantic labelling or pose estimation.
For these reasons, we decided to investigate how, to date, information about the depth of an image is collected and prepared in order to be used in a concrete way for different purposes. First of all we would like to underline that there are three main approaches to get the depth of a scene, that are using sensors, stereo images and monocular images. Different types of sensors can be used, such as rgbd cameras (i.e. Microsoft Kinect) for indoor scenarios, or LIDAR sensors for outdoor environments [1]. The convenience in their use lies in the fact that the information is obtained directly without any dependence on the scenario. However, the measurements obtained are scattered especially in outdoor situations, moreover, it must also be considered that this type of hardware is not so cheap and portable if we think of an application on a small robot or on an autonomous drone.
This is why the study of the other two approaches is of great importance for the scientific community. Regarding stereo images, the depth estimation is obtained by exploiting geometric constraints on camera motion and planarity, however, the main problem with this approach is that the accuracy is strongly limited by all the camera set-ups and the assumptions made to retrieve the data.
Moving on to monocular RGB images, the well-know issue of this procedure is that estimating the depth from a single image is an ill-posed problem. This means that, given a single image, an infinite number of possible world scenes may have produced it, therefore it could be infeasible to correctly estimate the global scale of the scene for that particular image. In order to tackle this fact, different techniques were developed, however, during the last years, the evolution of the Deep Learning field with the introduction of convolutional neural networks was a game-changer for computer vision in general, obtaining also interesting results for this specific task by using single images, having the groundtruth depth information stored pixel-wise [2].
The main factors that pushed us towards this approach, in addition to the results already obtained from previous research ([3], [4], [5]), are, first of all the fact of using a geometric-independent paradigm, moreover, the hardware for monocular images is more portable and cheaper than Lidar sensors and, finally, most of the datasets available about images are RGB single images datasets. With regard to this last point, we must highlight the fact that, although a large amount of data is available, these datasets are not enough to obtain domain independence of the scene, which means that if a model has been trained for example on a particular urban scenario, if it were tested on a scenario never seen as a forest, the performance of the prediction can drop drastically.

1.1 Approach

In this project we decided to develop an end-to-end neural network to predict the depth map of single RGB images as input for outdoor scenes. In order to counteract the problem presented previously regarding the independence of the scene, we decided to train our model using artificial data, i.e. video sequences of a virtual world produced by a graphics engine. We will therefore try to understand the goodness of using only synthetic data. If we observe good results from using just virtual datasets, this would suggest that we could produce artificial data from different scenarios in order to obtain a model that is independent of the specific scene.
Concerning the problem of understanding the global scale of the scene for the single image, we will help our depth estimation by adding a semantic information of the scene. We will obtain this



(a) Model architecture ”depth-mancini”.



(b) Model architecture ”DS-Net”.

Figure 1: Qualitative representation of the network architectures for the ’depth-mancini’ and ’DS-Net’ models. In blue convolutional layers, green max pooling and orange deconvolutional layers

by building a joint model with two related tasks: a regression task (depth estimation) and a classification task (semantic segmentation). We decided to introduce the semantic segmentation task because we thought about how a person elaborates the depth of objects. It seemed natural to consider that a person is facilitated in understanding the depth of a certain area of space through the association of that area to a specific object. To be clearer, having information about the semantic level of an image makes an important contribution in defining the depth of a certain region of the pixels of that scene. We will therefore examine what is the impact of adding semantic guidelines during training on the performances of the depth estimation task on this specific setup.

The report will be structured as follows:

- The section ”Related work” is dedicated to the deepening of previous researches related to the topic under consideration, with a particular attention to works that have taken similar approaches to our proposed method regarding models with parallel tasks and the use of synthetic datasets.
- In the ”Methodology” section, the models implemented will be described in detail, followed by the presentation of the dataset used and the details regarding the training sessions of the models.
- The data obtained from the training sessions of the networks are shown in the section ”results”, then, we propose a discussion and comparison of the models used.
- Finally, the section ”conclusions and future work” contains final thoughts on the project and the next steps to continue this research.

2 Related work

Previous research has successfully demonstrated the effectiveness of using Convolutional Neural Networks for regression problems such as estimating the depth of single RGB image input ([2], [6]). However, most of them have produced models focused only on a single task, also refining the models with respect to the specific dataset. Obviously there is no lack of recent works in which networks have been produced capable of carrying out multiple tasks at the same time, demonstrating how these models get better performances if trained on multiple tasks related to each other rather than simply on the single task.
Eigen et al. [7] has developed a multi-task model capable of predicting depth, pixel-wise semantic segmentation and the surface normals with a single RGB image as input. The main concept proposed in [7] is to predict at the beginning a coarse global output, and then use three other networks to refine the prediction to higher resolution. The same architecture is applied to each of the three tasks with the only difference for the last part of the network for the specific outputs.
A. Mousavian et al. [8] used a different approach for predicting both depth and semantic segmentation with a single network. The first layers of the model are shared for both tasks, this part is a multi-scale CNN which extracts shared features from the input image. Following, there are two separate branches, one for each task, where for the semantic branch it is used a Conditional Random Field to produce the final probabilities of semantic labels. Their work successfully showed that semantic segmentation and depth estimation could indeed share the features representation of the single image, boosting performances of both tasks when trained together.
A summary of the two methods just presented can be found in the work of P. Wang et al. [9], in which a first convolutional neural network is used, having as outputs the predictions of depth and semantic segmentation with respect to the global layout of the scene. After that, the image is divided into several regions and a second CNN produces the joint outputs on the local regions. Finally, global and regional outputs are both passed as inputs to a hierarchical CRF that infers the final results. The performance obtained suggested again that when these two tasks interact each other during training sessions, the network provides more accurate outputs then a CNN trained solely on one task.
The research discussed so far has always been based on using data obtained from real scenes, such as the KITTI dataset for outdoor scenes [1] and the NYU dataset [10] for indoor situations. However, research into the development of synthetic data for computer vision tasks is active and several papers have applied this to depth prediction, including Mancini et al. [11], in

which a virtual dataset was created with respect to two opposite scenarios, an urban situation and a forest. The main objective was to show the behaviors of models trained in different scenarios to better understand the problem of scene independence, showing performances on real data of models prepared on synthetic datasets.
Our proposed model takes its cue from the network developed in [11] for depth estimation; we add a branch dedicated to semantic segmentation, in order to understand and verify the effectiveness of the joint training of these two tasks even in a context where only artificial data are used.

3 Methodology

3.1 Models description

3.1.1 Depth-Mancini model

In order to have a yardstick to compare with our model, we decided to reproduce the proposed network in [11] which from now on will be called "depth-mancini". The network structure in question is an end-to-end network in encoder-decoder fashion. The encoder section is featured by the use of the feature extraction part of the VGG19 net [12] (pruned of its fully-connected layers). Therefore we find 16 convolutional layers all characterized by filters with a small kernel size (3x3), having a padding such as to preserve the spatial resolution of the input. Finally, among some of the 16 convolutional layers there are five max-pooling layers for the spatial pooling. The encoder weights are initialized with values of the VGG19 model trained to classify images on the Imagenet dataset [13], after which, this whole section is fine-tuned.
Moving on to the decoder part, it is composed of four deconvolutional layers for up-projections, followed by a convolutional layer having as a result an output with the same input resolution. The weights of this section, after several network optimization tests, were initialized using the uniform Xavier method [14].

3.1.2 DS-Net model

After this brief overview of the "depth-mancini" network, we continue with our proposed multi-task model, which we will call "DS-Net" during this report. The concept behind this network is simple, starting from the "depth-mancini" net, after the common encoder section there are two branches, one dedicated to the task for depth estimation, the second specialized in semantic segmentation. Consequently, both tasks share the feature extraction section.
The new branch used for semantics follows the same structure as the decoder already presented, with the only difference in the final part where the convolutional layer produces an output on 30 channels, each for each class for the semantic task. This output is then permuted to be passed as input of the softmax activation function, in order to obtain for each pixel the probability of belonging to one of the 30 classes.
During the training session two losses are calculated, the logarithmic root mean square error for depth estimation and the cross-entropy loss for semantic segmentation. These errors are then added together to produce a final loss from which the back-propagation starts on the whole model. For this reason, the feature extraction part is optimized with respect to both tasks (more details on the losses in the section dedicated to the description of the training). In figure 1, we reported qualitative representations of the two architectures.

3.2 Dataset description

As discussed previously, one of the main objectives of this project is to test the effectiveness of using synthetic data to learn how to predict the depth of a single RGB image. Consequently, the search for suitable datasets for this purpose has received particular attention. Therefore, after analyzing several proposals ([15], [16]), our choice has fallen on the Virtual Kitti [17], mainly because both groundtruths for depth and semantic segmentation are available in this dataset.

The Virtual Kitti dataset is composed of 50 video sequences from 5 different scenarios, so for each scenario we have 10 variants of the specific sequence, which are: camera rotated 15 degrees left / right, 30 degrees left / right, original sequence, presence of fog, during morning, overcast, raining and during sunset, for a total of 21,260 images. For each of these sequences, groundtruth is available for the following tasks: depth estimation, semantic segmentation, object detection, multi-object tracking and optical flow.
Specifically, the input images are 8-bit 3 channels RGB images of size 1242 x 375, while for the groundtruths of our interest we have that the depth groundtruth is stored in 16-bit single channel image, therefore with a 0 - 65535 range of values. For the semantic segmentation we have at our disposal qualitative outputs of RGB images with colors for classes and .txt files with the map between pixel color and the corresponding class.

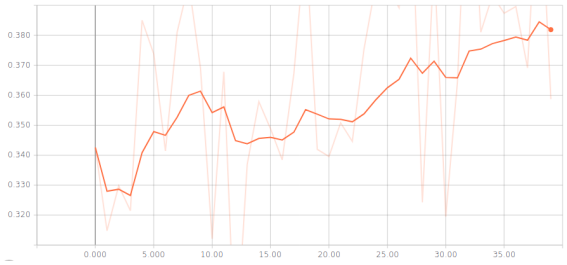
3.3 Training configuration

3.3.1 Data pre-processing

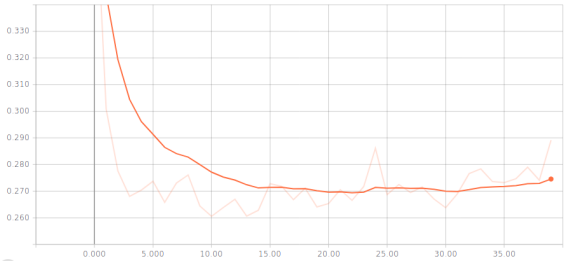
Input images before being fed into the network are rescaled to the dimension (160 x 256) in order to avoid memory issues and then normalized in the (0 - 1) range.

As far as depth groundtruth is concerned, since the values are saved in a 16 bit channel, particular attention must be given to modifying these data, since the risk of losing information is high. The final solution adopted is as follows: the groundtruths are resized with the same size of input, then we divide the depth values by 100.0, after this we clamp the values with a maximum value of 40.0. This is because, as suggested in the readMe file of the dataset, the depth values represent meters distance from 0 to 655.35. Therefore, after this pre-process, our values can represent a maximum depth of 40 meters.

Regarding the semantic segmentation, the ideal format for groundtruth are the so-called Label Images, i.e. images where each pixel has as a value the identifier of the class it belongs to. Given the particular format of groundtruth for semantic segmentation in the Virtual Kitti dataset, we had to convert the available qualitative images into label images using the class identifiers provided in the .txt files. Finally, the label images have been resized to the desired size.

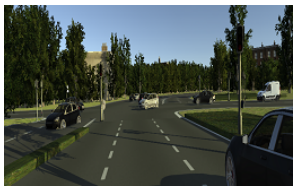


(a) Learning curve validation loss "depth-mancini".

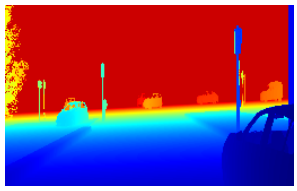


(b) Learning curve validation loss "DS-Net".

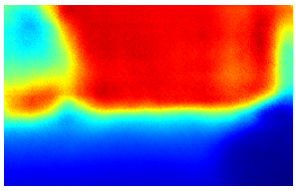
Figure 2: Validation losses learning curves comparison - The two plots show the behavior of the logRMSE loss for both models in the range of 40 epochs.



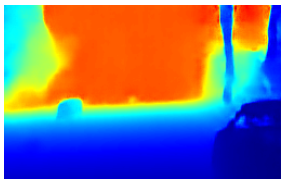
(a) input image



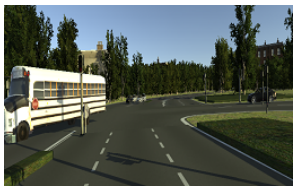
(b) colormap depth groundtruth



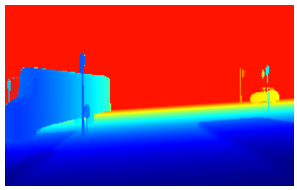
(c) DS-Net. LogRMSE: 0.2773



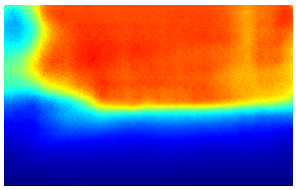
(d) 'd-mancini'. LogRMSE: 0.2933



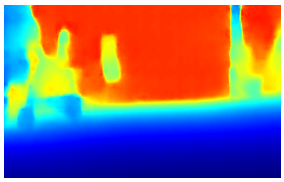
(e) input image



(f) colormap depth groundtruth



(g) DS-Net. LogRMSE: 0.2108



(h) 'd-mancini'. LogRMSE: 0.2501

Figure 3: Qualitative results on the test sequence in the Virtual Kitti Dataset

3.3.2 Training parameters specification

For the training sessions we used the Adam optimizer with a learning rate of 10e-5. We trained for 40 epochs with a batch size of 32, using as training set the scenarios ('0001', '0018', '0020'), as validation set the scenario '0006' and as test set the scenario '0002'.
The training sessions took 26 hours for the 'depth-mancini' model and 35 hours for the 'DS-net' model on single GPU Nvidia Tesla K40.

3.3.3 Losses

As mentioned earlier, the logarithmic Root Mean Square Error (LogRMSE) is optimized for the depth estimation task, while the cross-entropy loss is used for the semantic classification of the pixels. Below is a brief description of the two losses:

- LogRMSE: Let us consider we have the depth groundtruth Y and the predicted output Y^*. The LogRMSE loss is then defined as:

$$LogRMSE(Y, Y^*) = \sqrt{\frac{1}{T}\sum_{Y \in B} ||log(y_i) - log(y_i^*)||^2} \quad (1)$$

  where y_i is a single pixel of the groundtruth, y_i^* the corresponding pixel of the prediction and B is the set of the batch concerned.
  We decided to use this specific loss because we want our model to pay more attention to the objects closest to the camera. Having our output a range of values equal to (0.0 - 40.0), the use of LogRMSE allows us to exploit the non-linearity of the log function in order to give more weight to errors made on pixels that represent short distances.

- Cross-entropy: the output coming from the branch of semantic segmentation is produced by a softmax activation function, having as a consequence the number of channels equal to the number of classes. The cross-entropy loss is therefore:

$$crossEntropy(C, C^*) = -\frac{1}{n}\sum_i C_i log(C_i^*) \quad (2)$$

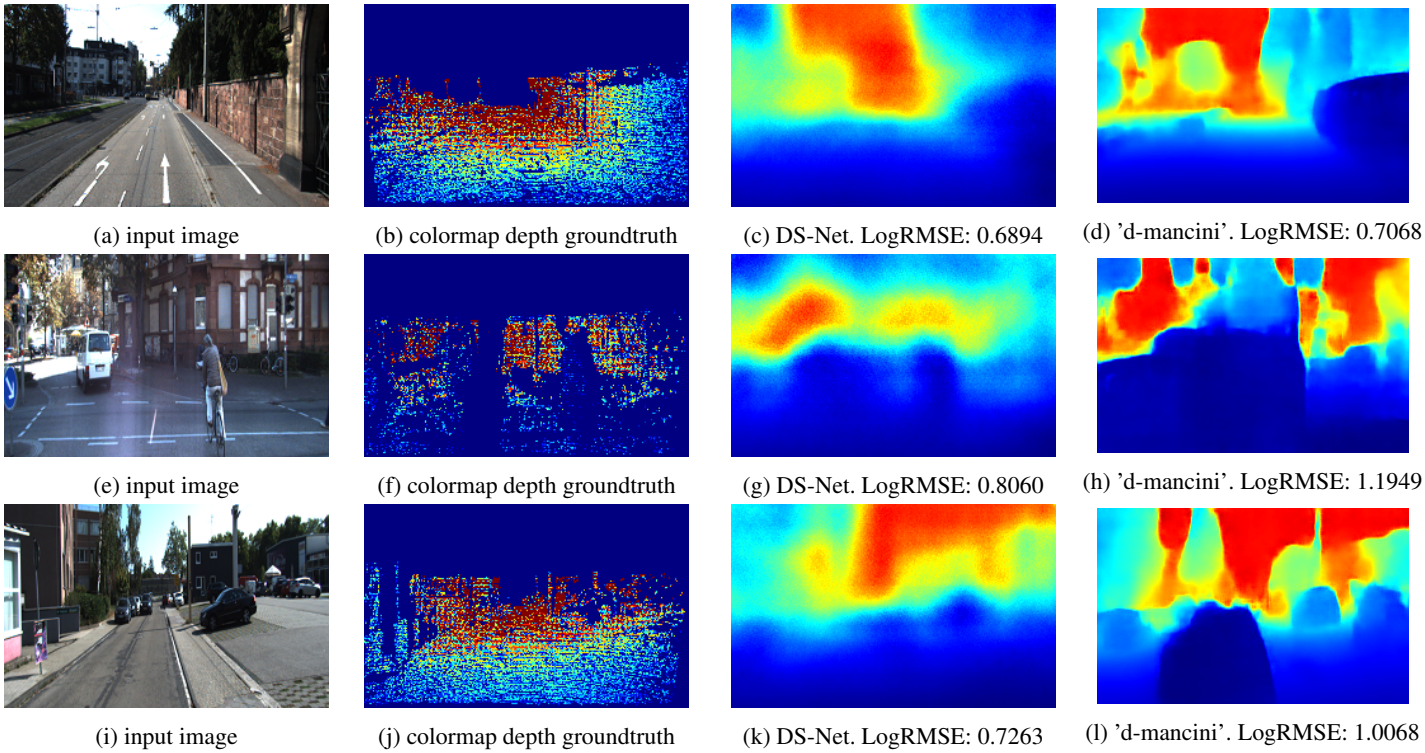
  where $C_i^* = e^{z_i} / \sum_c e^{z_{i,c}}$ is the class prediction at pixel i from the output z of the semantic branch.

4 Results

4.0.1 Learning curves

Figure 2 shows the learning curves of the loss with respect to the validation set for both the 'depth-mancini' and 'DS-net' models calculated at the end of each epoch.

As you can see, the validation loss of the 'depth-mancini' not only has a strong variance, but its behavior suggests a symptom of overfitting. On the other hand, our 'DS-net' model has a much more stable behavior on the validation loss, touching values 17.8% lower than the 'depth-mancini' loss minima.



(a) input image (b) colormap depth groundtruth (c) DS-Net. LogRMSE: 0.6894 (d) 'd-mancini'. LogRMSE: 0.7068

(e) input image (f) colormap depth groundtruth (g) DS-Net. LogRMSE: 0.8060 (h) 'd-mancini'. LogRMSE: 1.1949

(i) input image (j) colormap depth groundtruth (k) DS-Net. LogRMSE: 0.7263 (l) 'd-mancini'. LogRMSE: 1.0068

Figure 4: Qualitative results on the KITTI Dataset (real data)

From the results obtained with these learning curves we can formulate two aspects, namely, the joint training of the two tasks has made it possible to effectively reduce the loss of the depth branch; secondly, the addition of semantic information has contributed also as a natural regularizer of the depth task as suggested in the work of Wang et al. [9]

4.0.2 Evaluating on the test set

After discussing the learning curves of the two models we are going to show the performance of both networks on the test set. First of all we specify that the metrics used for the comparison are the LogRMSE and the Root Mean Square Error (eq. 3)

$$RMSE(Y, Y^*) = \sqrt{\frac{1}{T}\sum_{Y \in B} ||y_i - y_i^*||^2} \tag{3}$$

where again y_i is a pixel of the groundtruth, y_i^* the corresponding pixel of the predicted output and B is the set of the specific batch.

As it was possible to suppose from the trend of the validation losses, our 'DS-net' has gotten better results in both the metrics (table 1), underlining once again how the optimization of two correlated tasks can bring benefit to both, offering a concrete improvement in the ability of the model to generalize respect to the training data.

4.0.3 Test set qualitative results

As you can see from the set of images shown in figure 3, at a quality level we have some interesting results. Surprisingly, the 'depth-mancini' model, despite having shown unstable behaviour with regard to validation loss, demonstrates that it has effectively learned the shape of cars and, also, it is sometimes able to recognize thin objects such as tree trunks and poles (compare figures 3.c and 3.d).
On the other hand, the 'DS-net' shows to be more regular respect to scene change, where the 'depth-mancini' fails by recognizing shapes of objects that are not actually present or above the maximum limit of 40 meters, especially tree branches (compare figures 3.g and 3.h). This behavior is probably caused by the fact that the 'depth-mancini' has strongly overfitted the dataset and consequently it reproposes patterns learned during the training (tree branches, vehicles) even when these objects are not present, because, even if the test set is a scenario never seen, it has similarities compared to the sequences of the training set.

Table 1: Comparison on the metrics results for both models on the test set of the Virtual Kitti Dataset.

model	LogRMSE metric	RMSE metric
depth-mancini	0.27115	7.46425
DS-Net	**0.23945**	**6.50301**

4.0.4 Evaluating on KITTI dataset

After having carried out the necessary tests with respect to the Virtual Kitti Dataset, as the last step we decided to test the performance of both models on a dataset containing data derived from reality. For this purpose it was natural to choose the KITTI dataset [1] as the candidate, since Virtual Kitti is directly inspired by this dataset, as well as being one of the most used benchmarks for evaluating models on outdoor scene understanding tasks. As a sequence we used the data available in the folder 'val-selection-

Table 2: Comparison on metrics results on the KITTI dataset

model	LogRMSE metric	RMSE metric
depth-mancini	0.80617	11.97442
DS-Net	**0.76492**	**11.64267**

cropped' downloadable from the official site, a set of 1000 frames in urban scenario.
From the results shown in table 2, we can once again see that our joint model produces better results than the 'depth-mancini' model, even on real figures that have never been seen before.

4.0.5 Qualitative results KITTI dataset

For the qualitative results related to this test on real data, it is even more evident how 'depth-mancini' network suffers in its performance respect to scenes strongly different from those within the training set. If we focus on image 4.e, we notice the same behavior found in the qualitative results on the Virtual Kitti test set, where 'depth-mancini' produces patterns such as cars and trees even in situations of total absence of these objects (see also image 4.l). On the other hand, "DS-Net", while offering worse results compared to the ones on the Virtual Kitti test set, still it predicts values closer to the real scene in a regular manner.

5 Conclusion and Future Work

In this work, we have studied and explored a different approach based on Deep Learning in order to estimate the depth of a scene in outdoor environments represented by single RGB images, using only synthetic data. We therefore proposed an architecture based on the simultaneous execution of two scene understanding tasks, depth estimation and semantic segmentation, in order to investigate the effects on the performance of depth prediction caused by the addition of the scene's semantic information. The results obtained on both real and artificial data clearly showed us how the combined training on two related tasks causes a concrete improvement in the behavior of the network, making the model more robust in generalizing its predictions.
Finally, with regard to the effectiveness of the use of synthetic data, we have seen a drop in performance once the model has been tested on real datasets. We have therefore concluded that the artificial dataset used is not yet fully mature, especially with respect to the differentiation between proposed scenarios, changes of environment and lack of details within the sequences. As a future work, we aim to continue exploring the feasibility of using synthetic data for scene understanding tasks, focusing on how to make semantic information more explicit during the training.

References

[1] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[2] D. Eigen, C. Puhrsch, and R. Fergus, "Prediction from a single image using a multi-scale deep network," in *Proc. Conf. Neural Information Processing Systems (NIPS)*, vol. 2, no. 3, 2014, p. 4.

[3] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 239–248.

[4] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *European Conference on Computer Vision*. Springer, 2016, pp. 740–756.

[5] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2024–2039, 2016.

[6] Y. Kuznietsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6647–6655.

[7] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2650–2658.

[8] A. Mousavian, H. Pirsiavash, and J. Košecká, "Joint semantic segmentation and depth estimation with deep convolutional networks," in *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016, pp. 611–619.

[9] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille, "Towards unified depth and semantic prediction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2800–2809.

[10] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[11] M. Mancini, G. Costante, P. Valigi, T. A. Ciarfuglia, J. Delmerico, and D. Scaramuzza, "Toward domain independence for learning-based monocular depth estimation," *IEEE Robotics and Automation Letters*, vol. 2, no. 3, pp. 1778–1785, 2017.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255.

[14] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[16] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "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)*, June 2016, pp. 3234–3243.

[17] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," *arXiv preprint arXiv:1605.06457*, 2016.

[18] T. Dharmasiri, A. Spek, and T. Drummond, "Joint prediction of depths, normals and surface curvature from rgb images using cnns," *arXiv preprint arXiv:1706.07593*, 2017.

[19] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[20] O. H. Jafari, O. Groth, A. Kirillov, M. Y. Yang, and C. Rother, "Analyzing modular cnn architectures for joint depth prediction and semantic segmentation," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4620–4627.

[21] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1253–1260.